Changing the Conduct of Science in the Information Age

Summary Report of Workshop Held on November 12, 2010

National Science Foundation

June 28, 2011

Disclaimer: Any opinions, findings, conclusions, or recommendations expressed in this material are those of the workshop participants and do not necessarily reflect the views of the National Science Foundation.

Contents

Executive Summary

New digital technologies are transforming the practice of science. Science is now increasingly computational, data-intensive, and collaborative because digital technologies provide new ways for scientists to both create scientific information, and to communicate, replicate, and reuse scientific knowledge and data. These same technologies are creating important opportunities for international funding agencies to promote scientific collaboration and to foster the replication and reuse of scientific information.

The U.S. National Science Foundation (NSF) held a workshop titled "*Changing the Conduct of Science in the Information Age*" on November 12, 2010, to promote international cooperation in such policy areas as the promotion of data access, the development of technical solutions for open data platforms, and attribution for research contributions. This report describes the discussions, findings, and suggestions generated by the distinguished group of international workshop participants.

Participants identified a number of key findings with respect to data access. They noted that the primary social barriers to data access include insufficient intellectual property rights, the difficulty of documenting data for reuse, and problems associated with protecting confidentiality and privacy. They also noted that a host of technical issues must be addressed, including data control, security, long-term preservation, and stewardship. Participants agreed that scientific information should be broadly defined to include both data and code, and that knowledge sharing encompasses a variety of modes and methods. They noted that scientific attribution is critical to establishing trust in the research community and thus promoting knowledge access.

Workshop participants outlined a vision for the future that includes a framework for openness and international standards for data and knowledge; reliable and unique identifiers for individual researchers, organizations, and publications to create linkages between publications and their appropriate data; continuous investment for data preservation and access; and formal and informal training of students, researchers, and funding agency personnel.

There was a strong consensus that this vision could be achieved with the help of a concerted, collaborative effort by international funding agencies to:

(1) Establish a system of persistent identifiers for researchers and their outputs;
(2) Develop national and international pilot projects that compare different technical solutions for establishing and maintaining open data platforms, fostering the replication of scientific research, and ensuring attribution for the intellectual contributions of researchers; and
(3) Foster formal and informal training to develop scientists' skills in knowledge and data access, as well as data analysis.

1. Introduction

New digital technologies are transforming the practice of science. Science is now increasingly computational, data-intensive, and collaborative because digital technologies provide new ways for scientists to both create scientific information, and to communicate, replicate, and reuse scientific knowledge and data. Two key elements of this transformation are access to data and access to knowledge. Digital technology can make data openly accessible to scientists, reducing data-management burdens, formalizing generalizable and replicable science, and enabling new kinds of data-driven science. Digital technology can similarly facilitate the dissemination and transmission of knowledge by making information widely available electronically.

Institutional and professional barriers limit both data and knowledge access, however. For example, the costs associated with data access, including storage, documentation, and dissemination are not uniformly supported. Further, the current system of scientific attribution does not capture the complexity of contributions to scientific knowledge.

International funding agencies have an important opportunity to change policies to reduce these barriers. They could use digital technologies to promote scientific collaboration, and to foster the replication and reuse of scientific information, thereby changing the conduct of science. They could identify technical solutions for developing and maintaining open data platforms to promote collaboration and cooperation, foster the replication of scientific research, and ensure attribution for the intellectual contributions of researchers (National Science Foundation 2010b).[1]

To examine how these barriers might be overcome by international funding agencies and organizations supporting research, the U.S. National Science Foundation (NSF) held a workshop titled "*Changing the Conduct of Science in the Information Age"* on November 12, 2010. The workshop brought together members of the research community, computer and information scientists, as well as behavioral and social scientists, to identify guiding principles and approaches that could help inform organizations that fund research, scientific research organizations, and publishing houses.

The workshop placed questions into three categories:

- *Technical constructs*—What are the most important digital technologies that could be used to facilitate access to data and knowledge? To what extent is progress already being made, and how can progress be accelerated? What role might the private sector play in facilitating change?

[1] References given in parentheses are listed at the end of this report.

- *Social constructs*—What incentives are necessary to engage scientists in making data accessible and shared with the broader community? What are the appropriate business models necessary to promote connecting publications to data? How might private-sector participants be engaged in the effort? What are the social barriers to adopting and using unique researcher numbers?

- *The Pragmatic Experience*—What lessons have been learned from Brazil's experience with the Lattes platform? What opportunities are possible as a result of the establishment of the ORCID (Open Researcher and Contributor ID) project? What can be learned from data preservation, libraries and other coordinated data and publication efforts? What can be learned from domain-specific successes?

2. Data Access

Access to data generated by the "data deluge" is crucial.[2] Research reproducibility is critical (Hirsh 2010; Donoho 2010; Donoho et al. 2009), as "[r]eplicability is a hallmark of science" (Börner 2010), but research is only reproducible if the underlying data are accessible (Stodden 2010) and reliable. The challenge is daunting: one workshop speaker noted that the scientific community now generates more data each year than the entire sum of data produced in all prior years combined (Seidel 2010). Much data are inaccessible because of the dramatic increase in the amount of "information which is 'off the records' of science, not available to peer reviewers, [and] in many cases not even recorded in formal lab notebooks or laboratory information management systems" (Pfeiffenberger 2010). A recent NSF/Office of Cyberinfrastructure (OCI) Grand Challenges Task Force Report identified reproducibility of computational results as an

***Exemplar*: Sloan Digital Sky Survey (SSDS)**
The SDSS is a map of the universe that was compiled from 1991 to 2008. It has generated 850 million web hits in 9 years by 1,000,000 distinct users (globally there are only 15,000 professional astronomers). SDSS tops the astronomy citation list and has delivered more than 100 billion rows of data. It has facilitated both remote collaborations and discoveries by amateur scientists (http://www.sdss.org/).

[2] Recently held workshops and reports devoted to data access include the European Commission's High level Expert Group on Scientific Data, *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*, October 2010, available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf; Yale Law School, *Data and Code Sharing Roundtable*, November 21, 2009. See also http://www.law.yale.edu/intellectuallife/codesharing.htm, and "A Special Report on Managing Information: Data, Data Everywhere," *The Economist*, February 25, 2010, available at http://www.economist.com/node/15557443.

imperative challenge for the computational sciences (National Science Foundation 2010), and a roundtable at Yale University published a declaration urging reproducible computational research through the sharing of data and code (Yale Law School Roundtable on Data and Code Sharing 2010).

"Information overload" is a related challenge, where the frequency or volume of available data overwhelms the ability of an individual or organization to usefully process, classify, manage, or analyze them (Elias 2010). Much is inaccessible due to technical difficulties accessing the proliferation of different types of data available to researchers, including numerical arrays and experimental results (Stodden 2010).

Participants noted that funding agencies should act due to abundant evidence that data access is valuable in advancing science (Raddick and Szalay 2010; Donoho et al. 2009). This includes remote collaborations among scientists, such as those using data available through the Sloan Digital Sky Survey (see Exemplar) (Raddick and Szalay 2010; Neylon 2010a), European Supersites for Atmospheric Aerosol Research (EUSAAR), or biodiversity research (Wood 2010). A cyber community has collaborated on infrastructure development to advance biodiversity research. The European Space Agency validates satellite Earth observation data, CERN (The European Organization for Nuclear Research) enables the grids for e-science, and the Global Biodiversity Information Facility (GBIF) and Encyclopedia of Life (EOL) facilitate data access (Raddick and Szalay 2010). Data accessibility has also helped citizen science flourish (Raddick and Szalay 2010; Hirsh 2010; Stodden 2010), and it has contributed to the concept of "collective intelligence," which "seeks to understand [the] new ways in which people collaborate and create outcomes that are integrally about large groups of participating individuals, as much as they are about the new technologies that underlie them" (Hirsh 2010). Academic research may also be enriched by greater access to the abundance of data already being collected in the public and private sectors (German Data Forum (RatSWD) 2010), provided that privacy and proprietary rights are protected.

The workshop participants discussed the social and technical challenges associated with promoting data access.

2.1. Social Challenges

Workshop participants agreed that the primary social barriers associated with data access include insufficient intellectual property rights, the difficulty of documenting data for reuse, and the problems associated with protecting confidentiality and privacy.

There was agreement that academic institutions do not completely recognize the ownership and intellectual property rights relating to data production and sharing. There are also legal constraints: current copyright and other intellectual property laws in many nations present legal barriers to fully sharing data, articles, papers, methodologies, and code.

Workshop participants also noted that any comprehensive data-access plan must resolve the tension between confidentiality and openness (Schutz 2010). As the European Union has acknowledged, "[i]nnovation is important in today's society, but should not go at the expense of people's fundamental right to privacy"(EU NewsBrief 2010). While the German Data Forum has recognized the enormous research potential of allowing access to official census data and other sources of public data, it has also emphasized respect for individual privacy and the need to protect individually identifiable data (German Data Forum (RatSWD) 2010). Legal, ethical, and administrative restrictions on the reuse of data sets containing personally identifiable information, such as income, health, and criminal records (Elias 2010), are intended to safeguard human subject privacy, ensure subject consent for the use of personal data, or provide stewardship. These restrictions are often applied to data sets gathered and maintained by national statistical offices and agencies.

Participants made a number of suggestions for addressing social challenges related to data access. "Persuasive" incentives, such as attribution or linking data sets to subsequent publications, are needed to encourage researchers to give the wider research community access to their data (Pfeiffenberger 2010). Moreover, it is important to "value the publication of data (and software) as potentially equivalent to articles about conclusions, methods, instrumentation, models, algorithms and whatever is considered a legitimate object of publication" (Pfeiffenberger 2010). Published data could serve as an assessment and certification of quality, much as the publication of a peer-reviewed academic article represents the vetting of an argument or concept, and allow data sets to become part of "the scientific record" (Pfeiffenberger 2010). In addition, researchers that have openly provided their data to others should be recognized through attribution in any subsequent publication that makes use of the data (Trasande and Hannay 2010).

***Exemplar*: Earth System Science Data** The Data Publishing Journal provides quality assessments for data sets that reside in permanent repositories. The journal maps peer-review criteria from text to data (http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html).

One option for resolving the conflicts between reproducibility and copyright law is the development of an Open Research License (ORL) to "encourage researchers to create fully reproducible research by allowing [them] to capture more of the credit for facilitating and expanding scientific understanding, while promoting the ideal of reproducible research" (Stodden 2008).

Participants agreed that training is critical to maintaining open data systems, whether informal, formal, or through discussions and research practices. Training could take place at the graduate and post-doc levels, potentially instilling good habits at an early career stage. Colleges and universities could require graduate students and post docs to write papers using only publicly available data and publish them in open-source journals [Schutz; Trasande].[3] But without advisors setting an example by using open data, students would not see it as a laudable practice [Sauermann]. It is also important to develop a method for crediting informal training and learning to provide incentives to perpetuate the education of successive generations of researchers [Trasande].

***Exemplar*: Permanent Access to the Records of Science in Europe (PARSE) Insight**

The EU is funding PARSE, is a 2-year project focused on the preservation of digital information in science over time and ensuring that it "is accessible, usable and understandable in the future." The goal is to create a roadmap for facilitating continuous access to scientific data (http://www.parse-insight.eu/).

One potential method for enhancing access to data while maintaining human subject privacy is the use of "virtual safe settings—systems through which authorised and authenticated users can gain remote access to data on individuals or organisations whilst preventing copying of data and minimising the potential for abuse of access privileges" (Elias 2010).

2.2. Technical Issues

Workshop participants noted that increasing access to research data involves solving a host of technical issues, including data control, security, long-term data preservation, and stewardship.

Data storage should be planned from the start of any data-sharing enterprise (Wood 2010), and access control, archive security, and protection of confidential data should also be considered during the planning process (Schutz 2010). In this sense, the Permanent Access to the Records of Science in Europe Insight project (see Exemplar) serves as a potential model for scientific data infrastructure (Wood 2010). Providing useful access to the magnitude of research data that is continually being created is a primary challenge of any effort to create and harmonize a global scientific data infrastructure. One participant pointed out that scope presents the greatest barrier, asserting, "petabytes are easy, exabytes are hard" [Hirsh]. Not only must data be collected and stored, they must also be retrievable in discrete sets that can easily be reused by researchers.

Several options for addressing technical issues related to data access were discussed by participants. An important conceptual framework for creating a user-friendly scientific

[3] Last names in brackets refer to participants who made comments during the workshop. *See* Appendix C for participant biographies.

data infrastructure may include a "Knowledge Organisation System" that provides a consistent means of describing science, maintains an overview of the interrelationship between various areas of scientific knowledge, and presents an extract of the connections between past scientific knowledge and the emergence of new scientific knowledge from current and future research (Lambe 2010). Another possible model is a Reproducible Research System (RRS), which has two parts: (1) a Reproducible Research Environment (RRE) for computation work, which "provides computational tools together with the ability to automatically track the provenance of data, analyses, and results, and to package them (or pointers to persistent versions of them) for redistribution," and (2) a Reproducible Research Publisher (RRP), such as standard word-processing software or other documentation-preparation system, that links to the RRE. This facilitates readers' abilities to reproduce the analysis, as well as "extend it with the document itself by changing parameters, data, filters, and so on" (Pfeiffenberger 2010).

"Data...needs to be accessible by anyone, from anywhere, at anytime" (Viegas 2010). This requires the creation of taxonomies of scientific data to enable the cataloguing, tagging, and parsing of data sets for automated recall. Currently, a number of scientific data infrastructure systems use different and incompatible data identifiers, inhibiting data sharing and reuse (Fenner 2010). Another obstacle to the creation of a useful scientific data sharing infrastructure is the issue of interoperability—ensuring that researchers can easily reuse data sets originating in any country. To overcome this issue, international data standards and taxonomies must be explored by the scientific research community. This will likely first occur in individual disciplines, then in interdisciplinary conversations.

Standardized identification schemes, such as Altman and King's Universal Numerical Fingerprint implemented at the Dataverse Network at Harvard University (Altman and King 2007); metadata standards like MIAME for microarray gene expression (Trasande and Hannay 2010; Fenner 2010); and Digital Object Identifiers (DOIs) for any physical or digital manifestation, including text, audio, images, and software (Fenner 2010), can aid in categorizing and managing data and data sources and increase interoperability and ease of use. However, for these schemes to be successful, scientists must be encouraged and given incentives to routinely use the standards to annotate their data, a process that will be aided by the development of better and easier to use software tools (Trasande and Hannay 2010).

Not only do all students need to be trained to utilize open data, but individuals need to be formally educated as "data scientists" [Wood]. A new cohort of computational scientists who can manage the integration of data sets from disparate sources is essential [Aragão; Santos].

2.3. Role of Funding Agencies

Workshop participants suggested a number of ways in which funding agencies worldwide can significantly improve data access. For example, they can provide incentives for data sharing by publishing openness rankings [Evans], or using curriculum grants or similar measures to encourage informal and formal training of students on the importance of open data [Seidel]. Agencies can promote interoperability through international initiatives to develop persistent digital research data infrastructure.

Current efforts include, the U.S. National Science and Technology Council Interagency Working Group on Digital Data's recommendation that all U.S. federal agencies "promote a data management planning process for projects that generate scientific data for preservation" (Office of Science and Technology Policy 2009). NSF has complied with this call to action and changed the implementation of its long-standing data policy[4] by requiring that beginning in January 2011, all proposals include a "Data Management Plan" that describes:

- The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
- The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
- Policies and provisions for reuse, redistribution, and the production of derivatives; and
- Plans for archiving (National Science Foundation 2011).

Other funding agencies have also taken an active role in ensuring data access. For example, the Alliance of German Science Organisations published the June 2008 "Digital Information Initiative" designed "to equip scientists and academics with the information and infrastructures best suited to facilitate their scientific work" (Lauer 2010).[5] The United Kingdom's Economic and Social Research Council website (http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx) also notes:

[4] "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data...created or gathered in the course of work under NSF grants" (National Science Foundation 2010a).

[5] *See* also Deutsche Forschungs Gemeinschaft, http://www.dfg.de/en/research_funding/programmes/infrastructure/lis/digital_information/alliance_initiative/index.html.

> [T]hose ESRC grant applicants who plan to generate data are responsible for preparing and submitting data management and sharing plans for their research projects as an integral part of the application. It is then a responsibility of the award holder to incorporate data management and sharing as an indivisible part of the research project to increase the potential for data to be shared. We require that the data must be made available for preparation for [reuse] and/or archiving with the ESRC data service providers within three months of the end of the award otherwise we will withhold the final payment.

3. Knowledge Access

Sharing knowledge about scientific discoveries is a foundation of modern science, but workshop participants noted that funding agencies need to understand that knowledge, and therefore knowledge sharing, should be broadly defined to encompass both data and code. They also noted that knowledge sharing takes many forms and should be encouraged, including traditional academic journal publishing as well as other mechanisms such as discussion forums, recommendations, wikis, file-sharing sites, blogs, and microblogs (Trasande and Hannay 2010). However, substantial barriers to knowledge access persist despite mandates to promote sharing. For example, in spite of the embrace of Open Access publishing, the voluntary adoption rate by scientists has been low (around 15%–20%). Mandates have increased these numbers to around 70% for NIH-funded research and in institutions, such as Southampton or CERN, that have adopted these policies. Nevertheless, this means that even with mandatory participation, some 30% of research is not openly available (Fenner 2010).

3.1. Social Issues

Workshop participants agreed that attribution for new forms of scientific activity was critical to promoting knowledge access. Researchers will provide access to their work if they are given credit for their labor. Attribution for scholarly work requires the ability to uniquely identify both specific contributors to research and specific scientific contributions (Fenner 2010). Participants felt strongly that an author-identification system that transcends institutional, disciplinary, and national boundaries would help create a "clear and unambiguous scholarly record" of research activities associated with an individual and help provide unambiguous attribution for researcher contributions, whether they appear as publications, patents, or data sets (Office of Science and Technology Policy 2009; National Science Foundation 2011). An author-identification system would also allow for "microattribution" for research contributions not associated with a peer-reviewed journal publication (Credit Where Credit Is Due 2009). In the current system, a significant portion of scientific work remains unrecognized because there are no formal methods for providing attribution for this labor

(Pfeiffenberger 2010; Elias 2010; Lambe 2010; Neylon 2010b). Examples include the work of students, research assistants, and other non-author collaborators or the participants in large-scale scientific infrastructure projects. This renders much of the labor that makes science possible "invisible"—"the visible manifestations of science conceal an intricate social network of relationships, trust and perceived authority, underlying how science gets funded, how scientists decide to collaborate, and how new knowledge gets validated" (Lambe 2010).

With an international, unambiguous, and comprehensive attribution framework, data could be collected on the full range of research labor and outputs (Lane 2010), enabling a "wide range of derived metrics and rankings" (Trasande and Hannay 2010) that could be used to better understand the complexity of scientific labor and research. The Knowledge Organisation System described in Section 2.2 may help make "visible" the scientific labor previously left "invisible" by connecting formally recognized scientific outputs and metrics to the informal labor and social networks that support the generation, dissemination, and reuse of scientific knowledge.

Authentication, verification, quality assurance and control, and privacy provisions are critical to the success of a persistent author-identification system (Neylon 2010b). The systems must be able to handle duplication and redundancies, and should "not be affected by name changes, cultural difference in name order, inconsistent first-name abbreviations or the use of different alphabets" (Credit Where Credit Is Due 2009). There is a financial incentive for this as well: "In the current world ill-considered, non-transparent, and irreproducible metric systems will almost inevitably lead to legal claims" (Neylon 2010b).

At the heart of resolving this issue are establishing and authenticating unique researcher identifiers. To avoid misidentifications, access to individual researcher identifiers should be under the control of individual researchers (Neylon 2010b), and researchers should be required to authenticate their biographical and professional information (Trasande and Hannay 2010). A centralized database designed to populate grant and job applications, bio-sketches, or reports, which can otherwise be onerous and repetitive tasks (Evans 2010), would likely be an incentive for researchers to keep this information current. The Lattes Platform, a research database, adopts a related approach, requiring users to register before applying for government funding.

***Exemplar*: Open Researcher and Contributor ID (ORCID)**

Proposed in 2009, ORCID is a system of unique alphanumeric strings for each researcher. It is backed by 23 organizations, including Thomson Reuters, the British Library, and the Wellcome Trust (Credit Where Credit Is Due 2009). ORCID intends to create a central registry of unique identifiers linked to other author schemes (http://www.orcid.org/).

For a system based on unique researcher identifiers, all parties must trust that the identifier system is reliable, authentic, and immutable. Accordingly, the development of systems that are unable to generate a significant level of trust is "likely to limit and fragment any effort to coordinate, federate, or integrate differing identity solutions in the research space. Therefore interoperability of any developed system with the wider web must be a prime consideration" (Neylon 2010b). One possible method for establishing trust and ensuring proper attribution is encouraging the use of an identification system like ORCID (see Exemplar), and establishing publishing practices that make the "creation and capture [of unique identifiers] an integral part of the editorial process" (Trasande and Hannay 2010).

3.2. Technical Issues

The development of a persistent, trusted, ubiquitous, and interoperable centralized repository for housing the unique researcher identifiers may provide a "trusted broker" for promoting knowledge access and attribution (Trasande and Hannay 2010; Neylon 2010b). Currently, a number of identification tools exist or are under development, including ORCID, Vivo, Lattes Platform, Public Library of Science (PLoS), and PubMedCentral. For example, ORCID proposes to create a "central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes" (http://www.orcid.org/). Lattes, a fully developed researcher database that allows for verification (Aragão 2010; Lane 2010) has now been adopted in 17 countries in Latin America, Europe, and Africa [Aragão].

3.3. Role of Funding Agencies

Funding agencies worldwide can play a critical role in encouraging knowledge access and the implementation of an identification system to facilitate attribution. Agencies are uniquely positioned to require data and code sharing in publicly funded work, and they support the infrastructure and tools for data and code sharing. Participants felt that funding agencies should embrace the creation of identification systems and ensure their adoption by requiring registration as a prerequisite to applying for agency funding (Trasande and Hannay 2010). Participants also thought that agencies could support a research library coalition that would provide an international open-standard data set for bibliometric information for all published work worldwide (Conlon 2010).

4. Conclusions and Next Steps

At the conclusion of the workshop, participants agreed on a set of attributes of the "ideal" attribution landscape 5 years into the future [Greer]. It would include a framework of openness and international standards for data and knowledge; reliable and unique identifiers for each researcher, organization, publication, and the relationship to each other; a link between all publications and their appropriate data; continuous investment for data preservation and access; and formal and informal training of students, researchers, and personnel at funding agencies.

In white papers submitted before the workshop, and during presentations and in discussions at the workshop itself, participants identified a set of actions that would achieve this vision:

(1) Establish a system of persistent identifiers for both researchers and their outputs. The following specific suggestions were made about the characteristics of such a system:

- Create taxonomies of scientific data—Enable the cataloguing, tagging, and parsing of data sets for automated recall.
- Create incentives—Encourage and offer incentives to researchers to routinely use the standardized identification schemes to annotate their data, a process that will be aided by the further development of software tools. Provide researchers with incentives to encourage them to make data sets available to the wider research community through the development and use of attribution systems. Help ensure that data sets are linked to subsequent publications and other research outputs, further aiding attribution and the reproducibility of research. Publish data and code to facilitate assessment and certification of quality and allow data sets to become part of the citable "scientific record."
- Create independent standards—Establish federally funded platforms for data and code sharing that are independent of institutions and individual researchers, and use standards of unique identification for citation and version control.
- Create a legal framework—Develop an Open Research License (ORL) to resolve conflicts between reproducibility and copyright law.
- Create a registration mechanism—Encourage the development, implementation, and use of standardized identification systems to facilitate attribution by requiring system registration as a prerequisite to applying for agency funding.

(2) Develop national and international pilot projects that compare different technical solutions for developing and maintaining open data platforms, fostering the replication of scientific research, and ensuring attribution for the intellectual contributions of researchers.

(3) Foster formal and informal training to ensure that open data and knowledge systems are maintained.

Workshop participants agreed that engaging in these efforts will provide opportunities to work across counterpart funding agencies to encourage international cooperation and the dissemination of knowledge and data.

Appendix A: Workshop Background

> It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study.
>
> —NSB Report: "Long-lived Data Collections: Enabling Research and Education in the 21st Century," September 2005.

Workshop Motivation

Digital technology offers the potential for fundamental change in the conduct of science for two reasons:

- Data access—Digital technology could make data openly accessible to scientists, reducing data-management burdens, formalizing generalizable and replicable science, and enabling new kinds of data-driven science.

- Knowledge access—Digital technology could facilitate the dissemination and transmission of knowledge by making information widely available electronically.

These technologies can transform science by enabling a broad range of scientists to create and transmit knowledge, educators to impart that knowledge to future generations, and decision-makers to make well-informed policy based on sound and reproducible research. But institutional and social barriers that exist limit the acceptance of these transformational technologies by the scientific community. It is likely that these barriers could be reduced by concerted and informed efforts by scientific funding agencies and international organizations.

An informed approach to digital technology can be fostered by advancing understanding of the technical, behavioral, and social factors conducive to its widespread adoption. These factors might include the role of economic incentives, human capital, social networks, and (most obviously) scientific attribution. Lessons can also be learned from practical experience. Fields as disparate as biotechnology, geosciences, and astronomy have been transformed by both data and knowledge access. Several broad-based initiatives, such as ORCID, the Brazilian Lattes Platform, and the VIVO project, have promoted widespread access to knowledge. Both the research and practical experiences should help identify new approaches that are neither nation specific nor domain specific—indeed, that can be used in a cooperative international effort to help foster the adoption and use of digital technologies.

Objective

The goal of this workshop is to combine the expertise of computer and information scientists with those of behavioral and social scientists to identify guiding principles and approaches that can help inform organizations that fund research, scientific research organizations, and publishing houses. Specific questions that could be answered include the following:

> *Technical constructs*—What are the most important digital technologies that could be used to facilitate data and knowledge access? To what extent is progress being made already, and how can progress be accelerated? What role might the private sector play in bringing about change?
>
> *Social constructs*—What incentives are necessary to engage scientists in making data accessible and shared with the broader community? What are the appropriate business models necessary to promote connecting publications to data? How might private-sector participants be engaged in the effort? What are the social barriers to adopting and using unique researcher numbers?
>
> *The Pragmatic Experience*—What lessons have been learned from ORCID and the Brazilian experience? What can we learn from data preservation and libraries and other coordinated data and publication efforts? What can we learn from domain-specific successes?

Workshop Structure

The workshop is intended to be small and informal in nature, and the discussion will be focused on addressing the motivating questions. To achieve that goal, the workshop will have four separate sessions. Selected participants will be asked to start the dialogue by providing a short opening commentary on their experience with data and knowledge access, focusing on describing the technical and social challenges and how what was learned might help inform international efforts. All workshop participants are asked to participate in a discussion of opportunities and constraints. The discussions should focus on relating enabling aspects of the technologies to incentives for changing conduct, with the goal of identifying approaches that are related to the technological solutions. It is particularly important that the discussion address social, technical, and institutional challenges, using specific examples from previous experience. The moderator will provide a synthesis of the discussion and identify useful approaches for funding agencies.

To facilitate planning, all workshop participants for the sessions are asked to produce, in advance, a brief document summarizing their views about new approaches that could be used to foster the adoption and use of digital technologies. Although these views might be informed by either theory or practice, providing a list of relevant literature and examples would be extremely helpful. All participants will receive the briefing documents in advance, and it is expected that these documents will inform and lead the discussion at the workshop.

Appendix B: Workshop Structure

A diverse group of international research scientists, computer and information scientists, and behavioral and social scientists was invited to participate in the workshop. Invitations were also extended to representatives from funding agencies worldwide and publishers likely to participate as partners in open-access data initiatives. This provided a broad base of expertise encompassing social, professional, and technical issues related to data and knowledge access.

Workshop Wiki

NSF created a wiki to facilitate sharing information and ideas both before and after the workshop.[6] Before the workshop, background readings and participant biographies were provided, and each participant was asked to draft a white paper. The white paper was to be a brief document summarizing the participant's views on what approaches could be used to encourage adopting information and communication technologies to enable open access to data and data sharing. All workshop participants received briefing documents in advance of the workshop, with the expectation that these documents would inform the workshop discussions. In the spirit of openness, following the workshop, the presentations were posted and comments were solicited on each.

Sessions

By design, the workshop was small and informal in nature to focus discussion on the workshop's motivating questions. To achieve this goal, the workshop had four separate sessions:

1. Data Access—Digital technology and scientific communities.
2. Data Access—Digital technology and multiple scientific communities.
3. Knowledge Access—The role of scientific attribution.
4. Knowledge Access—The role of funding agencies.

Selected participants were asked to initiate and moderate the sessions by providing brief commentary on their experience with data and knowledge access, particularly technical and professional challenges and how the lessons learned from these might help inform international efforts to address open access to data more broadly. All workshop participants were encouraged to participate in the ensuing discussion of

[6] A "wiki" (Hawaiian for "fast") is a page or collection of interlinked Web pages designed to enable collaborative websites. Anyone who accesses the wiki is able to contribute or modify content using a simplified markup language.

opportunities and constraints, with the goal to develop “ideas, exemplars and concrete recommendations” (Seidel 2010). The discussions focused on identifying approaches that addressed both the potential of ICTs to enable data sharing and open access to data and the incentives for researchers, agencies, institutions, and publishing houses to facilitate this. Participants also addressed the technical and professional challenges to data sharing, drawing on specific examples from prior experience. At the end of each session, the moderator was asked to provide a synthesis of the discussion and identify useful approaches for funding agencies. For the purposes of this report, we have separated the workshop into two categories: (1) data access and (2) knowledge access and attribution.

Appendix C: Participant and Observer Biographies

Workshop Participants

Dr. Carlos Aragão, Professor of Physics, Universidade Federal do Rio de Janeiro. Dr. Aragão obtained a B.S. in Physics in 1973 and an M.S. in Physics in 1976 at the Pontificia Universidade Católica do Rio de Janeiro (PUC/RJ). He obtained a Ph.D., also in Physics, from Princeton University in 1980. Dr. Aragão is a Professor of Physics at the Universidade Federal do Rio de Janeiro (UFRJ). He is a Member of the Brazilian Academy of Sciences and was awarded the Brazilian National Order of Scientific Merit Medal in 1998.

Dr. Shenda Baker, Professor of Chemistry, Harvey Mudd College. Dr. Baker received a B.S. in Chemistry and French from Grinnell College in 1985. In 1991, she received a Ph.D. in Physical Chemistry from the California Institute of Technology. Dr. Baker is a Professor of Chemistry at Harvey Mudd College in California, where she became the Clare Boothe Luce Assistant Professor of Chemistry in 1993. In 1996, Dr. Baker received an NSF CAREER Award, the DOE Young Scientists and Engineers Award, and the Presidential Early Career Award for Scientists and Engineers. She is on the Executive Committee for the National Neutron Scattering Society of America and on the Advisory Board for the Office of Cyberinfrastructure at NSF.

Dr. Katy Börner, Victor H. Yngve Professor of Information Science, School of Library and Information Science, Indiana University. Dr. Börner received an M.S. in Electrical Engineering from the University of Technology in Leipzig in 1991 and a Ph.D. in Computer Science from the University of Kaiserslautern in 1997. She is the Victor H. Yngve Professor of Information Science at the School of Library and Information Science, Adjunct Professor in the School of Informatics, Core Faculty of Cognitive Science, Research Affiliate of the Biocomplexity Institute, Fellow of the Center for Research on Learning and Technology, Member of the Advanced Visualization Laboratory, and Founding Director of the Cyberinfrastructure for Network Science Center at Indiana University. She also serves as a curator of the Places & Spaces: Mapping Science exhibit.

Dr. Sayeed Choudhury, Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center, Sheridan Libraries, Johns Hopkins University. Dr. Choudhury holds graduate degrees in Civil Engineering and Systems Analysis and Economics from Johns Hopkins University. He is the Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center at the Sheridan Libraries of Johns Hopkins University. Dr. Choudhury is also the Director of Operations for the Institute of Data Intensive Engineering and Science (IDIES) at Johns Hopkins, a Lecturer in the Department of Computer Science at Johns Hopkins, a Research Fellow at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, and a Senior Presidential Fellow with the Council on Library and Information Resources.

Dr. Elaine Collier, Assistant Director for Clinical Research, Division for Clinical Research Resources, National Center for Research Resources, U.S National Institutes of Health.

Dr. Michael Conlon, Director of Data Infrastructure, University of Florida. Dr. Conlon received a Bachelor's degree from Bucknell University and a Ph.D. in Statistics from the University of Florida. He is Director of Data Infrastructure for the University of Florida. Dr. Conlon is also Research

Associate Professor of Biostatistics at the University of Florida and co-principal investigator of INVEST—the International Verapamil/Trandolapril Study, a randomized trial of 22,000 patients at 860 sites in 14 countries. The trial is conducted entirely online using Internet-based software designed and developed by a team under Dr. Conlon's direction. The software approach has been patented and licensed to MarCon Global Data Solutions, a company he co-founded along with Dr. Ronald Marks.

Dr. Peter Elias, Professor, Institute for Employment Research, University of Warwick. Dr. Elias received a Bachelor's degree in Chemistry before undertaking his doctoral studies in applied labor economics at the University of California, Berkeley. He is a Professor at the Institute for Employment Research, University of Warwick. He is a Fellow of the Royal Statistical Society, and has been the Strategic Advisor for Data Resources to the UK Economic and Social Research Council (ESRC) since 2004.

Dr. James Evans, Assistant Professor of Sociology, University of Chicago. Dr. Evans received a B.A. from Brigham Young University in 1994 and an M.A. and Ph.D. from Stanford University in 1999 and 2004, respectively. He is Assistant Professor of Sociology at the University of Chicago and a member of the Committee on the Conceptual and Historical Studies of Science. He is also a Fellow at the Computation Institute.

Dr. Martin Fenner, Hannover Medical School Cancer Center and ORCID Board of Directors. Dr. Fenner studied medicine in Berlin, and he did further training in basic and clinical oncology, including a postdoctoral fellowship in Boston. He now works at the Hannover Medical School Cancer Center. He writes the weblog Gobbledygook (http://blogs.plos.org/mfenner), is on the board of directors of ORCID (Open Researcher & Contributor ID; http://www.orcid.org/), and helps organize the Science Online London (http://www.scienceonlinelondon.org/) conference.

Dr. Chris Greer, Assistant Director for Information Technology Research and Development, U.S. Office of Science and Technology Policy. Dr. Greer received a Ph.D. in Biochemistry from the University of California, Berkeley, and did his postdoctoral work at the California Institute of Technology before teaching at the University of California, Irvine, in the Department of Biological Chemistry. Dr. Greer is Assistant Director for Information Technology Research and Development in the White House Office of Science and Technology Policy (OSTP). He also serves as co-chair of the Interagency Working Group on Digital Data, which has been charged by the Committee on Science of the National Science and Technology Council with developing and promoting implementation of strategic frameworks for digital scientific data preservation and access.

Dr. Tony Hey, Corporate Vice President, External Research Division, Microsoft Research. Dr. Hey received a Bachelor's degree in Physics and a Ph.D. in Theoretical Physics from Oxford University. He is corporate vice president of the External Research Division of Microsoft Research, where he is responsible for worldwide external research (ER) collaboration in Microsoft Research. Dr. Hey is a fellow of the U.K. Royal Academy of Engineering and has served on several national committees in the U.K., including committees of the U.K. Department of Trade and Industry and the Office of Science and Technology. He is a fellow of the British Computer Society, the Institute of Engineering and Technology, the Institute of Physics, and the U.S. American Association for the Advancement of Science (AAAS).

Dr. Haym Hirsh, Professor of Computer Science, Rutgers University. Dr. Hirsh received his B.S. in Mathematics-Computer Science from UCLA and his M.S. and Ph.D. in Computer Science from Stanford University. He is Professor and past-Chair of Computer Science at Rutgers University. From 2006 to 2010 he served as Director of the Division of Information and Intelligent Systems at the U.S. National Science Foundation, managing an organization responsible for over $500 million of research grants in Computer and Information Science and Engineering. He has also held visiting positions at Bar-Ilan University, CMU, MIT, and the University of Zurich.

Mr. Patrick Lambe, Adjunct Professor in Knowledge Management, Hong Kong Polytechnic University, and Principal Consultant, Straits Knowledge. Mr. Lambe studied at Oxford University. He is now based in Singapore and is an expert in knowledge management, knowledge organisation systems, and taxonomies. He has weblogs at http://www.greenchameleon.com and http://www.organisingknowledge.com. Patrick is currently working with the National Science Foundation Division of Science Resources Statistics on taxonomy management and development projects to support that division's mission.

Dr. Gerhard Lauer, Professor of German Literature, Georg-August-Universität Göttingen. Dr. Lauer received his Ph.D. in 1992. He is professor of German Literature at Georg-August-Universität Göttingen in Germany. Dr. Lauer is also co-editor of the *Journal of Literary Theory,* a member of the German Research Council commission for electronic publishing, an advisory board member of the Open Access Publishing in European Networks (OAPEN; http://www.oapen.org), a board member for the European Science Foundation Bibliometric Database Scoping Project (2008-2010), coordinator for German-Israeli academic exchange, and head of the TransCoop-programme committee of the Alexander von Humboldt-Stiftung.

Ms. Ruth Lee, Director, Research Councils UK Office in the United States. Ms. Lee received a B.A. from the University of Sheffield and a Master of Education from the University of Manchester. She is the Director of the U.S. office for the Research Councils UK, the primary public body in the UK charged with funding research and supporting the next generation of researchers.

Dr. David Lipman, Director, National Center for Biotechnology Information, National Library of Medicine, U.S. National Institutes of Health. Dr. Lipman earned a B.A. in Biology from Brown University in 1976 and an M.D. from the State University of New York, Buffalo, in 1980. He is the Director of the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine within the National Institutes of Health (NIH). He was appointed as NCBI's first Director in 1989, shortly after Congress created the Center in 1988, and has overseen its growth into one of the most heavily used resources in the world for the search and retrieval of biomedical information, with about 2 million users each day. He is a member of the National Academy of Sciences, the Institute of Medicine, and the American Academy of Arts and Sciences.

Mrs. Lucia Carvalho Pinto de Melo, President, Center for Strategic Studies and Management. Mrs. Carvalho received a Bachelor's degree in Chemical Engineering from Universidade Federal de Pernambuco in 1973; a Masters degree in Physics from the Universidade Federal de Pernambuco in 1976; a Masters degree in Energy and Environment at the University of California, Santa Barbara, in 1980; and a Masters degree from the Technology and Policy Program at the Massachusetts Institute of Technology in 1987. From 1990 to 1991 she led the Department of

Science and Technology of Pernambuco State, and she was President of the Foundation for Research of the State of Pernambuco (FACEPE) from 1995 to 1998. Mrs. Carvalho is currently the President of the Center for Strategic Studies and Management (CGEE).

Dr. August Muench, Astronomer, Harvard-Smithsonian Center for Astrophysics. Dr. Muench received a B.S. in Physics from the Georgia Institute of Technology in 1995 and a Ph.D. in Astronomy from the University of Florida in 2002. He is an astronomer at the Harvard-Smithsonian Center for Astrophysics and a member of the Seamless Astronomy team at Harvard, which is a collective of projects funded by NASA, NSF, and Microsoft Research working to better integrate astronomy's numerous open-access data repositories, comprehensive literature catalogues, and diverse software tools.

Dr. Theodore Papazoglou, Policy Analyst, European Research Council. Dr. Papazoglou received a B.S. in Physics from the University of Crete in 1985 and a Ph.D. in Biomedical Engineering from the University of Southern California in 1989. After completing his postdoctoral training at the Laser Research Center at Cedars Sinai Medical Center in Los Angeles, he attained tenure at the Institute of Electronic Structure and Lasers of the Foundation for Research and Technology Hellas (FO.R.T.H.-IESL). In 2001, Dr. Papazoglou was detached to the European Commission and worked as a scientific officer at the *Marie Curie Fellowships* programme. At the end of 2003 he was recruited as temporary agent in the Directorate General for Research of the European Commission. He is now working in Unit A1 (Support to the ERC Scientific Council) of the ERC Executive Agency, and implementation of the ERC's Open Access strategy is among his duties.

Dr. Hans Pfeiffenberger, Leader, IT Infrastructure Department, Alfred Wegener Institute for Polar and Marine Research. Dr. Pfeiffenberger leads the IT infrastructure department of the Alfred Wegener Institute. He is a speaker for the Helmholtz Open Access working group and advises the Knowledge Exchange (DFG, JISC, SURF, and DEFF). In 2008, Dr. Pfeiffenberger was appointed representative of the Helmholtz Association to the Alliance on Permanent Access (APA). He is also chief editor of the journal *Earth System Science Data,* an innovative journal providing quality assurance to published data through peer review.

Dr. Marcio de Miranda Santos, Executive Director, Centre for Strategic Management and Studies in Science, Technology and Innovation. Dr. Santos earned an M.S. in Genetics and Plant Breeding and a Ph.D. in Biochemical Genetics. Dr Santos is the current Executive Director of the Centre for Strategic Management and Studies in Science, Technology and Innovation and Chair of the Board of Trustees of the Center of Reference on Environmental Information (CRIA). He has been a consultant to the FAO (Food and Agriculture Organization of the United Nations), the Interamerican Institute for Cooperation in Agriculture, Bioversity International (formerly the International Plant Genetic Resources Institute), and national governments on policies for the conservation and use of plant genetic resources for food and agriculture. Dr. Santos has represented Brazil in the FAO/Commission on Genetic Resources for Food and Agriculture and in the UNEP/CBD Conference of the Parties (COP).

Dr. Lorenza Saracco, Research Programme Officer, Research Infrastructure Unit, European Commission. Dr. Saracco is a research programme officer within the Research Infrastructure Unit of the European Commission. She joined the EC in 2003, and since then she has followed policies and projects dealing with data and ICT infrastructures. Her background is in computer science (her degree is from the University of Pisa, Italy), and before joining the Commission she worked for the Italian National Research Council (CNR), first as researcher in the area of conceptual modeling and management of Earth science data and then as policy officer dealing with the European Union research policies. As an EC officer she participated in various task forces and groups on data organisation and management.

Dr. Henry Sauermann, Assistant Professor, College of Management, Georgia Institute of Technology. Dr. Sauermann holds undergraduate degrees in Economics and Business Administration from the University of Potsdam, Germany, and a Ph.D. in Business Administration from Duke University. He is an Assistant Professor at the College of Management, Georgia Institute of Technology. One stream of Dr. Sauermann's research examines scientists' pecuniary and non-pecuniary motives and incentives and their effects on research productivity in industry as well as academia. A second stream of research focuses on scientific labor markets and on the career choices of junior scientists.

Dr. Bernard Schutz, Director, Max Planck Institute for Gravitational Physics, Albert Einstein Institute. Dr. Schutz earned a Ph.D. in physics from the California Institute of Technology. He is a Director of the Max Planck Institute for Gravitational Physics (Albert Einstein Institute) in Potsdam, Germany, and a member of the management team of the Max Planck Digital Library. He is also a Fellow of the American Physical Society and of the German Leopoldina Society, and he is a recipient of the Amaldi Gold Medal of the Italian Society for Gravitation. Dr. Schutz is active in science outreach and recently co-founded the Scienceface project (http://www.scienceface.org/), which is currently releasing a series of highly praised video interviews of leading scientists by a young interviewer with no scientific background; these are designed to make black-hole physics and astronomy accessible to young people.

Dr. Victoria Stodden, Assistant Professor, Department of Statistics, Columbia University. Dr. Stodden completed a Ph.D. in Statistics at Stanford University and a Masters in Legal Studies at Stanford Law School, where she created a new licensing structure for computational research. Previously, she had been a Postdoctoral Associate in Law and a Kauffman Fellow in Law at the Information Society Project at Yale Law School. She is an assistant professor of Statistics at Columbia University. She is currently co-chairing a working group on Communities and Virtual Organizations in the NSF Office of Cyberinfrastructure Task Force on Grand Challenge Communities. She is a Science Commons fellow, a member of the Sigma Xi scientific research society, and a member of the AAAS. Her website, which includes talks and publications, is http://www.stodden.net, and she occasionally blogs at http://blog.stodden.net.

Dr. Alex Szalay, Professor, Department of Physics and Astronomy, Johns Hopkins University. Dr. Szalay is the Alumni Centennial Professor of Astronomy and a professor in the Department of Computer Science at Johns Hopkins University. He is a cosmologist, working on the statistical measures of the spatial distribution of galaxies and galaxy formation. Dr. Szalay was born and educated in Hungary and has written over 450 papers in various scientific journals, covering areas from theoretical cosmology to observational astronomy, spatial statistics, and computer science. He is a Corresponding Member of the Hungarian Academy of Sciences and a Fellow of the American Academy of Arts and Sciences. Dr. Szalary received an Alexander Von Humboldt Award in Physical Sciences in 2004 and a Microsoft Award for Technical Computing in 2008.

Dr. Caitlin Trasande, Resident Scientist and Analyst, Nature Publishing Group. Dr. Trasande received a B.A. in Philosophy from St. John's College and a Ph.D. in Neuroscience from the University of Chicago in 2004. After serving as a postdoctoral scholar at Yale School of Medicine and Mount Sinai School of Medicine, she joined Nature Publishing Group (NPG) as a resident scientist and analyst in its technology unit, working on digital content management, special-content collections, and content search. From 2007 to 2009 she spearheaded development of a science metrics platform. In 2010 she and the science metrics project moved to a new digitally focused business unit within Macmillan (NPG's parent company), codename "Project Babbage."

Dr. Evelyne Viegas, Director, Microsoft Research. Dr. Viegas completed a Ph.D. in France. She is responsible for the Data Intelligence initiative at Microsoft Research in Redmond, WA. Before her present role, Dr. Viegas was a Technical Lead at Microsoft, delivering Natural Language Processing components to projects for MSN, Office, and Windows. Before Microsoft, and after completing her Ph.D. in France, she worked as a Principal Investigator at the Computing Research Laboratory in New Mexico on an ontology-based Machine Translation project.

Dr. Gert Wagner, Professor of Economics, Berlin University of Technology. Dr. Wagner received a Bachelor's degree in Economics in 1978 from the Johann Wolfgang Goethe University in Frankfurt, Germany, and a Ph.D. from the Berlin University of Technology in 1984. He is a professor of Economics at the Berlin University of Technology and a Max Planck Fellow at the MPI for Human Development in Berlin. Dr. Wagner is also Director of the German Socio-Economic Panel Study (SOEP) at DIW Berlin. He is chairman of the German Census Commission and German Council for Social and Economic Data, and he serves on the Advisory Board to Statistics Germany. He is a member of the Working Group on Social Sciences and Humanities of the European Strategy Forum for Research Infrastructures (ESFRI), the Founding Committee of the International Data Forum (IDF), and the Research Resources Board of ESRC/UK. He is research fellow of IZA and a Research Associate of CEPR. In 2007, Dr. Wagner was awarded the "Knight's Cross" of the Order of Merit of the Federal Republic of Germany.

Dr. John Wood, Professor, Imperial College London. Dr. Wood holds Doctoral degrees from Cambridge and Sheffield Universities. He is currently senior international relations adviser at Imperial College London. He will become the Secretary-General of the Association of Commonwealth Universities in July. Dr. Wood is a non-executive director of a number of companies, including Bio-Nano Consulting, and sits on the advisory board of the British Library. He is on the board of the Joint Information Services Committee, which is responsible for the UK academic computing network, and chairs its Support for Research Committee. He also chairs the European Commission's high-level group on the future management of scientific data. Dr. Wood was elected as a fellow of the Royal Academy of Engineering in 1999 and was made a Commander of the British Empire in 2007 for "services to science."

Workshop Participants from the U.S. National Science Foundation

Dr. Philip Bogden, Program Officer, Office of Cyberinfrastructure, U.S. National Science Foundation. Dr. Bogden holds a B.A. in Engineering and Applied Sciences from Harvard and a Ph.D. in Oceanography from Scripps Institution of Oceanography at the University of California, San Diego. He is currently the Program Director for the Office of Cyberinfrastructure at the

National Science Foundation. Dr. Bogden also holds a position as a Research Professor at the Center for Land-Sea Interactions at the University of New England. Before coming to the NSF, he was the CEO of GoMOOS, Inc., a private nonprofit organization with member institutions representing a broad array of stakeholders interested in ocean observations. GoMOOS manages a multi-institutional partnership that collects a wide variety of real-time ocean measurements. From 2003 to 2009, Dr. Bogden was also the Acting Director of the Southeastern Universities Research Association SURA Coastal Ocean Observing and Prediction (SCOOP) Program.

Dr. Myron Gutmann, Assistant Director, Directorate for Social, Behavioral & Economic Sciences, U.S. National Science Foundation. Dr. Gutmann holds a Ph.D. from Princeton University, and he is currently the Assistant Director of the Directorate for Social, Behavioral & Economic Sciences at the National Science Foundation. Before his appointment in 2009, Dr. Gutmann was the director of the Inter-University Consortium for Political and Social Research and a Research Professor at the Population Studies Center at the University of Michigan, where he was also a professor of history. He has a broad range of interests in interdisciplinary historical population studies relating population to agriculture, the environment, and health. He also studies ways that digital materials can be properly preserved and shared and how the confidentiality of research subjects can be protected when data about them is made available for secondary use.

Dr. Julia I. Lane, Program Director, Science of Science & Innovation Policy Program, U.S. National Science Foundation. Dr. Lane holds an undergraduate degree in Economics and Japanese from Massey University in New Zealand and an M.A. in Statistics and a Ph.D. in Economics from the University of Missouri-Columbia. She is the Program Director of the Science of Science & Innovation Policy program at the National Science Foundation. Her previous jobs included Senior Vice President and Director, Economics Department at NORC/University of Chicago; Director of the Employment Dynamics Program at the Urban Institute; Senior Research Fellow at the U.S. Census Bureau; and Assistant, Associate, and Full Professor at American University. Dr. Lane has organized over 30 national and international conferences, received several national awards, given keynote speeches all over the world, and served on a number of national and international advisory boards. She is one of the founders of the LEHD program at the Census Bureau, which is the first large-scale, linked employer-employee data set in the United States. She is also a fellow of the American Statistical Association.

Dr. Cora Marrett, Assistant Director, Directorate for Education and Human Resources, U.S. National Science Foundation. Dr. Marrett received a B.A. from Virginia Union University in 1963, an M.A. in 1965, and a Ph.D. in 1968 from the University of Wisconsin, Madison, all in Sociology. She served as University of Wisconsin’s senior vice president for academic affairs for 6 years before coming to NSF. Before her appointment at the UW System, Dr. Marrett served as senior vice chancellor for academic affairs and provost at the University of Massachusetts-Amherst for 4 years. She was a member of the UW-Madison faculty from 1974 to 1997, with appointments in sociology and Afro-American studies. Dr. Marrett advanced from associate professor to full professor and was associate chairperson of the Department of Sociology (1988–1991). She was affiliated with the Energy Analysis and Policy Program and the Wisconsin Center for Education Research. She received an honorary doctorate from Wake Forest University in 1996, and she was elected a fellow of the American Academy of Arts and Sciences in 1998 and the American

Association for the Advancement of Science in 1996. She is widely published in the field of sociology and has held a number of public and professional service positions.

Dr. Edward H. Seidel, Acting Assistant Director, Directorate for Mathematical & Physical Sciences, U.S. National Science Foundation. Dr. Seidel earned a Ph.D. from Yale University in Relativistic Astrophysics. He is Acting Assistant Director of the Mathematical and Physical Sciences Directorate at the National Science Foundation. Dr. Seidel is a physicist recognized for his work on numerical relativity and black holes, as well as in high-performance and grid computing. In June 2008, the National Science Foundation selected Seidel as its director for the Office of Cyberinfrastructure (OCI). On 1 September 2008, he began this position, in which he oversees advances in supercomputing, high-speed networking, data storage and software development on a national level. He has recently assumed the role of Acting Assistant Director for Mathematics and Physical Sciences at NSF.

Workshop Observers

Dr. Stefano Bertuzzi, Office of Science Policy, U.S. National Institutes of Health. Dr. Bertuzzi received a Ph.D. in Molecular Biotechnology at the Catholic University of Milan, Italy, and after postdoctoral training in the Laboratory of Molecular Neurobiology at the Salk Institute in San Diego, became a tenured Associate Professor at the Dulbecco Telethon Institute in Milan, Italy. Dr. Bertuzzi is responsible for return-on-investment analyses in the Office of Science Policy, Office of the NIH Director, U.S. Department of Health and Human Services. In this position, Dr. Bertuzzi advises the NIH Director on a wide range of health-science policy matters. He is the recipient of several NIH Director's awards, along with other national and international awards.

Dr. Rachel Bruce, Innovation Director for Digital Infrastructure, JISC. Dr. Bruce is the Innovation Director for Digital Infrastructure. She oversees innovation programs and activities that are funded by the Support for Research committee and the Infrastructure and Resources committee. These include a number of programs, for example digital preservation, management of research data, and geospatial infrastructure and resources. She is concerned with the updating of infrastructure for the creation, sharing, and managing of digital resources and related shared services, as well as the policy and practices required to improve their reuse and exploitation to enhance education and research.

Ms. Sarah Colon, Research Associate, Japan Science & Technology Agency. Ms. Colon has an undergraduate degree in Biochemistry from Cornell and Master's degrees in Advanced Japanese Studies from Sheffield University and International Economics and Public Policy from the Johns Hopkins University School of Advanced International Studies (SAIS). She is a research associate with the Japan Science and Technology Agency, an independently administered sub-agency of Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). She works at the liaison office in Washington, D.C., and follows and reports on U.S. science and technology trends for the headquarters in Tokyo.

Dr. Diane DiEuliis, Senior Policy Advisor, U.S. Office of Science and Technology Policy. Dr. DiEuliis received a Ph.D. in Biological Sciences at the University of Delaware. She then completed a research fellowship at the National Institutes of Health intramural research program in cellular neurobiology, focusing on the molecular and morphological features of neuronal cells. Following her laboratory research, Dr. DiEuliis became a program director at the National Institute of Neurological Disorders and Stroke, where she began managing the Alzheimer's and Parkinson's disease portfolio of research grants and programs. She developed several strategic research plans for Parkinson's disease, coordinating with the Department of Defense and Veteran's Administration programs, which helped to expand and diversify the field of federal research on Parkinson's. She now maintains many of these planning programs annually, and manages the Udall Centers program. Dr. DiEuliis is also working as a senior policy advisor in the President's Office of Science and Technology Policy. Her policy focus is within the life sciences, and she is the staff director for several subcommittees within the Committee on Science, including research business models, human subjects' research, and the science of science policy.

Dr. Amy Friedlander, Senior Advisor, Directorate for Social, Behavioral and Economic Sciences, U.S. National Science Foundation. Dr. Friedlander graduated from Vassar College, where she was elected to Phi Beta Kappa. She holds an M.A. and Ph.D. from Emory University and an M.S.L.I.S. from The Catholic University of America. She works with the Assistant Director for SBE to coordinate a strategic planning exercise to articulate the driving questions in the SBE sciences for the year 2020 and beyond. She also helps the directorate to develop cooperative work within NSF (CISE, Engineering, and OCI) and with other federal agencies (e.g., NEH, NARA, and Library of Congress). Dr. Friedlander is also Editor-in-Chief of the *ACM Journal on Computing and Cultural Heritage.* Before joining NSF in June 2010, she was Director of Programs at the Council on Library and Information Resources.

Dr. Daniel Goroff, Senior Policy Analyst, U.S. Office of Science and Technology Policy. Dr. Goroff earned his B.A. and M.A. degrees in Mathematics from Harvard University as a Borden Scholar, an M.Phil. in Economics at Cambridge University as a Churchill Scholar, a Masters in Mathematical Finance at Boston University, and a Ph.D. in Mathematics at Princeton University as a Danforth Fellow. On loan to the White House Office of Science and Technology Policy (OSTP), Dr. Goroff is a Program Director at the Alfred P. Sloan Foundation, working on science, technology, education, and economics. He is currently on leave from Harvey Mudd College in Claremont, California, where he is Professor of Mathematics and Economics and where he previously served as Vice President for Academic Affairs and Dean of the Faculty.

Dr. Neil Jacobs, Acting Programme Director, Information Environment, JISC. Dr. Jacobs is Acting Programme Director for Digital Infrastructure (Information Environment). He oversees a variety of projects and programs in the areas of access to and management of digital resources, including linked data and digital repositories, scholarly communications, and research information management. These cover issues of technical interoperability, cultural and organizational change, sustainability, and business models.

Mr. Kei Koizumi, Assistant Director for Federal Research and Development, U.S. Office of Science and Technology Policy. Mr. Koizumi received his M.A. from the Center for International Science, Technology, and Public Policy program at George Washington University, and he received his B.A. in Political Science and Economics from Boston University. He is a Fellow of the American Association for the Advancement of Science. He joined OSTP in February 2009 after serving on the Obama transition team as part of the Technology, Innovation and Government Reform Policy Working Group. Before joining OSTP, Koizumi served as the longtime Director of the R&D Budget and Policy Program at the American Association for the Advancement of Science

(AAAS). He is known as a leading authority on federal science and technology funding and budget issues and is a frequent speaker to public groups and to the press.

Dr. Tanu Malik, Research Associate, Computation Institute, University of Chicago. Dr. Malik earned an undergraduate degree from the Department of Civil Engineering at Indian Institute of Technology, Kanpu, and an M.S. and Ph.D. in Computer Science from Johns Hopkins University. Dr. Malik is a Research Associate with the Computation Institute (CI) at the University of Chicago. Her research interests are in issues relating to building large-scale data-management systems such as federating distributed data systems, replicating large databases, data approximation, data provenance, and data quality. A recurrent theme in her research is to reexamine the core principles of database technology in the light of new requirements emerging from scientific data. Her research has resulted in some innovative database technology for handling large amounts of distributed scientific data.

Ms. Jeri Metzger Mulrow, Senior Mathematical Statistician, Division of Science Resources Statistics, U.S. National Science Foundation. Ms. Mulrow holds a B.S. in Mathematics from Montana State University and an M.S. in Statistics from Colorado State University. She is Senior Mathematical Statistician in the Division of Science Resources Statistics at the National Science Foundation. Ms. Mulrow is currently the project leader of the SRS Taxonomy Project, senior advisor to the SRS Early Career Doctorates Survey, and lead author of the State Chapter for Science and Engineering Indictors 2012. She was named a fellow of the American Statistical Association (ASA) in 2010, is a member of the ASA Board, a senior member of the American Society for Quality, and a member of the American Association of Public Opinion Research. As a statistician in the federal statistical system, she is particularly interested in data quality, data usability, data visualization, data access, data sharing, and the role that taxonomy plays in all of it.

Dr. James Onken, Special Assistant to the Acting Deputy Director for Extramural Research, U.S. National Institutes of Health. Dr. Onken received an M.S. and Ph.D. in Psychology from Northwestern University and an M.P.H. with a concentration in Biostatistics from George Washington University. He is responsible for analyzing and presenting data on NIH research programs and research personnel for use in program evaluation and policy studies. He is also program manager for the NIH Research Portfolio Online Reporting Tool (RePORT) website (http://RePORT.nih.gov/), the RePORT Expenditures and Results (RePORTER) system, and the companion ExPORTER site, where users can download databases of NIH-funded projects and publications and patents citing support from NIH.

Dr. Walter Schaffer, Research Training Officer, Extramural Research Training and Career Development Programs, U.S. National Institutes of Health. Dr. Walter Schaffer is the NIH Research Training Officer responsible for the extramural research training and career development programs. He received a Ph.D. in Chemistry from the University of Texas at San Antonio in 1978, with a dissertation on oxidative metabolism in rat brains. He then served as a Staff and Senior Staff Fellow in the Lab of Metabolism at the National Institute of Alcohol Abuse and Alcoholism. In 1986 Dr. Schaffer began a career as a Research Training Officer. He is a Captain in the U.S. Public Health Service Commissioned Corps.

Dr. Mya Sjogren, Performance and Accountability Analyst, Office of Research and Development, U.S. Environmental Protection Agency. Dr. Sjogren works in the Environmental Protection Agency (EPA) Office of Research and Development (ORD). She led the performance and accountability team, which reviews the performance for the agency's research and development programs. She directed the development of ORD's stakeholder surveys and has contributed to internal and external evaluation efforts such as bibliometric analysis and organizational scorecards. She facilitated the EPA-sponsored NAS study on Evaluating Research Efficiency, the Board of Scientific Counselor reviews, and the internal pilot that assesses which EPA research is cited in regulatory decisions, including rules, guidance, and records of decision.

Dr. Michael Stebbins, Assistant Director, Biotechnology, U.S. Office of Science and Technology Policy. Dr. Stebbins received his B.S. at SUNY Stony Brook and Ph.D. in Genetics while working at Cold Spring Harbor Laboratory. He is the Assistant Director for Biotechnology at the White House Office of Science and Technology Policy. Before joining OSTP he was the Director of Biology Policy for the Federation of American Scientists. He is a co-founder of Scientists and Engineers for America and a former Adjunct Professor of Bioethics at University of Pennsylvania. He has worked as a Legislative Fellow for Senator Harry Reid and on policy issues at the National Human Genome Research Institute.

Dr. George Strawn, Chief Information Officer, U.S. National Science Foundation. Dr. Strawn has an undergraduate degree from Cornell College and holds a Ph.D. in Mathematics from Iowa State. Since 1991 he has been at the National Science Foundation, where he is currently the Chief Information Officer (CIO). He was Director of the Directorate for Computer and Information Science and Engineering (CISE) Division of Advanced Networking Infrastructure and Research and the NSFNET Program Director. Before working at NSF, Dr. Strawn was a computer science faculty member at Iowa State University, where he also held several administrative positions. From 1986 to 1995 he served as Director of the ISU Computation Center. Under his leadership, ISU became a charter member of the regional NSFNET network, MIDnet, and ISU created a thousand-workstation academic system based on an extension of the MIT Athena system. From 1983 to 1986 he served as Chair of the ISU Computer Science Department. Dr. Strawn currently serves as co-chair of both the interagency Large Scale Networking Working Group and the international Coordinating Committee for Intercontinental Research Networks. He served as co-chair of the interagency Federal Networking Council from 1995 to 1997. Dr. Strawn also has held several positions in the computer industry and has worked as an information technology consultant in both private industry and government.

Dr. Edmund (Ned) Talley, Program Director for Channels, Synapses and Circuits, National Institute of Neurological Disorders and Stroke, U.S. National Institutes of Health. In 2001, Dr. Talley received his Ph.D. from the University of Virginia (UVA), studying the physiology and pharmacology of motor neurons involved in respiration. After his Ph.D., he remained at UVA as a Research Assistant Professor. He initiated investigations into the CNS functions of two-pore-domain potassium channels, with an emphasis on their modulation by neurotransmitters and clinically important drugs. Dr. Talley joined the NINDS in 2005 as a Program Director for Channels, Synapses and Circuits. His program at the NINDS is focused on basic research in synaptic transmission and neuromodulation.

Appendix D: List of Background Papers

Blue Ribbon Task Force on Sustainable Digital Preservation and Access, "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information," National Science Foundation, February 2010, https://extwiki.nsf.gov/download/attachments/5799975/BRTF_Final_Report.pdf?version=1&modificationDate=1270471033207.

Choudhury, Sayeed, Benjamin Hobbs, and Mark Lorie, "A Framework for Evaluating Digital Library Services," *D-Lib Magazine* 8, July/August 2002, http://www.dlib.org/dlib/july02/choudhury/07choudhury.html.

European Union, "European Union Observes Data Protection Day," http://eurunion.org/emailcampaigns/preview.php?previewtype=html&nl=21&c=237&m=158&s=6b9351a6f2b8c87a0a5d1e223a2907d4.

European Union, "Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data," a final report of the High Level Expert Group on Scientific Data, A submission to the European Commission, October 2010.

German Data Forum (RatSWD), "RatSWD Working Paper Series No. 150: Recommendations for Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences," Federal Ministry of Education and Research, July 2010.

Greer, Chris, "NSTC Releases Strategy for Digital Scientific Data," 23 March 2009, http://www.dtic.mil/dtic/pdf/announcements/IWGDDRelease.doc&#124.

Hammond, Tony, Timo Hannay, and Ben Lund, "The Role of RSS in Science Publishing," *D-Lib Magazine* 10, December 2004, http://www.dlib.org/dlib/december04/hammond/12hammond.html.

Hsinchun Chen, Ronald N. Kostoff, Chaomei Chen, Jian Zhang, Michael S. Vogeley, Katy Börner, Nianli Ma, Russell J. Duhon, Angela Zoss, Venkat Srinivasan, Edward A. Fox, Christopher C. Yang, and Chih-Ping Wei, "AI and Global Science and Technology Assessment," *IEEE Intelligent Systems* 24, no. 4: 68–88, July/Aug. 2009, doi:10.1109/MIS.2009.68, http://www.computer.org/portal/web/csdl/doi?doc=doi/10.1109/MIS.2009.68.

Kahn, Robert E., and Jay Allen Sears, "A Brief Overview of the Digital Object Architecture," Corporation for National Initiatives, October 2003, https://extwiki.nsf.gov/download/attachments/5799975/OverviewDigitalObjectArchitecture.pdf?version=1&modificationDate=1270582345245.

Kahn, Robert E., and Robert Wilensky, "A Framework for Distributed Digital Object Services," *International Journal on Digital Libraries* 6 (2): 115–23, DOI 10.1007/s00799-005-0128-x, published online 13 March 2006, https://extwiki.nsf.gov/download/attachments/5799975/fulltext.pdf?version=1&modificationDate=1270582327918.

Lane, Julia, "Let's Make Science Metrics More Scientific," *Nature* 464: 488–89, 25 March 2010, doi: 10.1038/464488a, Published online 24 March 2010, http://www.nature.com/nature/journal/v464/n7288/full/464488a.html.

Lipman, David, "Ten Years of PubMed Central," *Columbia University Libraries Library News,* 26 January 2010, http://library.columbia.edu/news/exhibitions/2010/20100308_pubmed.html.

Mesirov, Jill P, "Accessible Reproducible Research," *Science* 327: 415–16, doi: 10.1126/science.1179653, 22 January 2010, https://extwiki.nsf.gov/download/attachments/5799975/Sciencemag+2009+01+23+Accessible+reproducible+research.pdf?version=1&modificationDate=1270470638792.

National Institutes of Health, "Research Portfolio Online Reporting Tools (RePORT)," U.S. Department of Health and Human Services, http://projectreporter.nih.gov/exporter/.

National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," National Science Foundation, September 2005.

Neylon, Cameron, "Science in the Open, An Openwetware Blog on the Challenges of Open and Connected Science: New Year, New Me," 23 January 2010, http://blog.openwetware.org/scienceintheopen.

Neylon, Cameron, and Shirley Wu, "Open Science: Tools, Approaches, and Implications," *Pacific Symposium on Biocomputing* 14: 540–44, 2009, http://psb.stanford.edu/psb-online/proceedings/psb09/workshop-opensci.pdf.

Office of Science and Technology Policy, "Harnessing the Power of Digital Data for Science and Society," Report of the Interagency Working Group on Digital Data to the National Science and Technology Council, January 2009, http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

Pullinger, John, and Gert G. Wagner, "RatSWD Working Paper Series No. 153: On the Respective Roles of National Libraries, National Archives and Research Data Centers in the Preservation of and Access to Research Data," German Data Forum, August 2010, https://extwiki.nsf.gov/download/attachments/5799975/RatSWD_WP_153.pdf?version=1&modificationDate=1289417297212.

Raddick, Jordan M., and Alexander S. Szalay, The Universe Online," *Science* 329: 1028–29, 27 August 2010, http://www.sciencemag.org/content/329/5995/1028.full.

Sauermann, Henry, and Roach, Michael, "The Price of Silence: Scientists: Trade-Offs between Pay and the Ability to Publish," 24 October 2010, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1696783.

Stodden, Victoria, "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright," http://www.stanford.edu/~vcs/papers/LFRSR12012008.pdf.

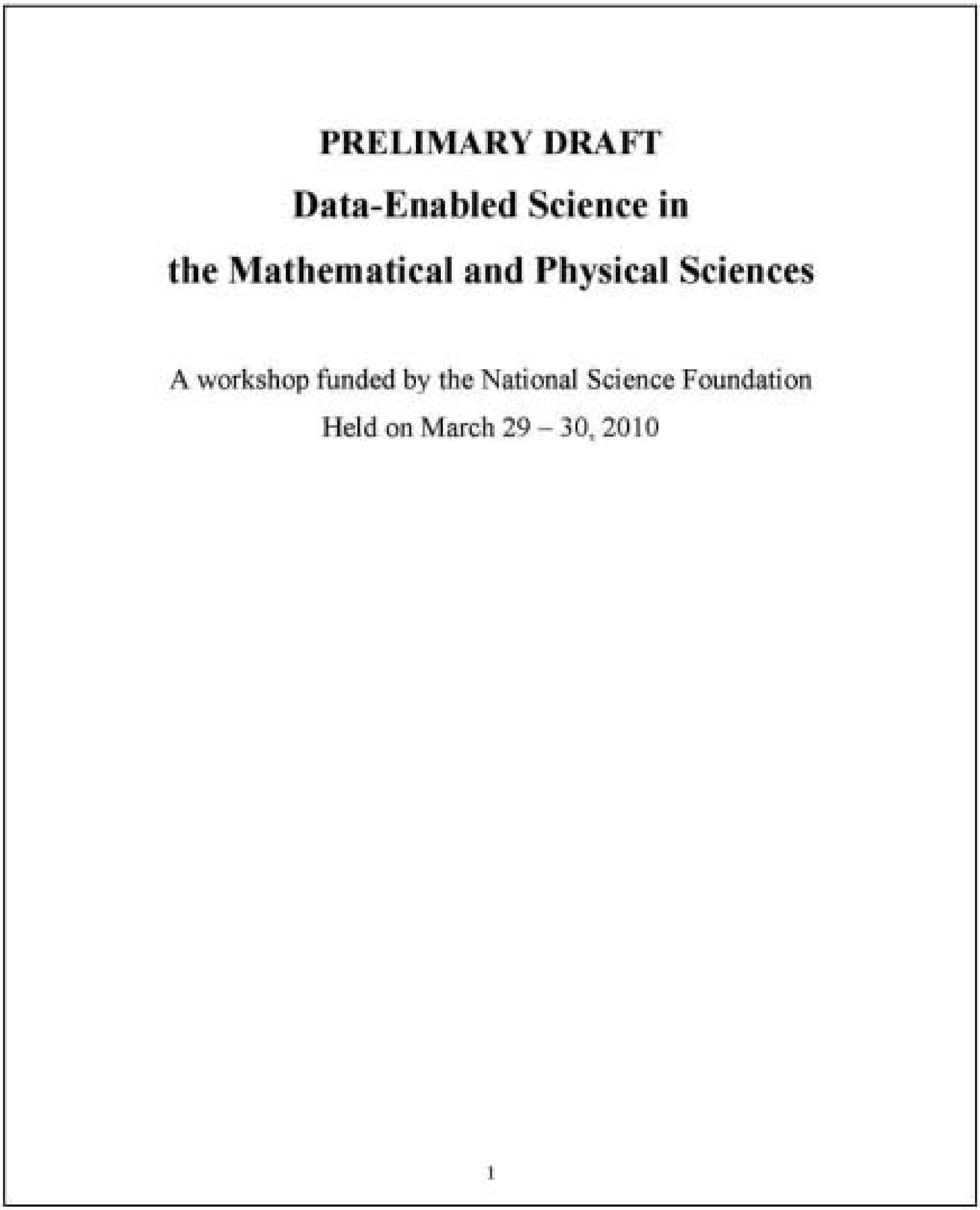
UK Data Forum, "A UK National Strategy for Data Resources for Social and Economic Research 2009–2012," http://www2.warwick.ac.uk/fac/soc/ier/publications/2009/nds_publication_sep09.pdf.

Appendix E: List of White Papers from November 12, 2010, Workshop

Author/Title
Baker, Shenda, et al., "Data-Enabled Science in the Mathematical and Physical Sciences: Workshop Report"
Börner, Katy, "Briefing Document for Changing the Conduct of Science in the Information Age"
Conlon, Mike, "The Objects of Science and Their Representation in eScience"
Elias, Peter, "Digital Technology and the Conduct of Scientific Research"
European Union, "Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data"
Evans, James, "Identification and the Complex System of Research"
Fenner, Martin, "White Paper for Changing the Conduct of Science in the Information Age"
Fenner, Martin, "Scientific Attribution Principles"
German Data Forum (RatSWD), "RatSWD Working Paper Series No. 150: Recommendations for Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences"
Hey, Tony, "Open Access, Open Data, Open Science"
Hirsh, Haym, "How Do You Cite a Crowd?"
Lambe, Patrick, "Changing the Conduct of Science in the Information Age: Discussion Points"
Lauer, Gerhard, "Changing the Conduct of Science in the Information Age: Focusing on Sharing Knowledge and Data"
National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century"
Office of Science and Technology Policy, "Harnessing the Power of Digital Data for Science and Society"
Papazoglou, Theodore, "IT-Based Approaches in Support of ERC's Mission to Support 'Frontier Research': First Experiences"
Pfeiffenberger, Hans, "Focusing on Social Constructs"
Sauermann, Henry, "Discussion Points for Session 3: Social Constructs; in Particular: Incentives"
Schutz, Bernard, "Data Access: Digital Technology and Scientific Communities"
Trasande, Caitlin, and Timo Hannay, "Changing the Conduct of Science: A Publisher's Perspective"
Viegas, Evelyne, "Data as an Enabler of Open Innovation: Challenges and Opportunities"

To access an extended version of this report that includes the workshop participants' white papers, go to: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=oise11003.

Baker, Shenda, et al., "Data-Enabled Science in the Mathematical and Physical Sciences: Workshop Report"



PRELIMARY DRAFT

Data-Enabled Science in the Mathematical and Physical Sciences

A workshop funded by the National Science Foundation
Held on March 29 – 30, 2010

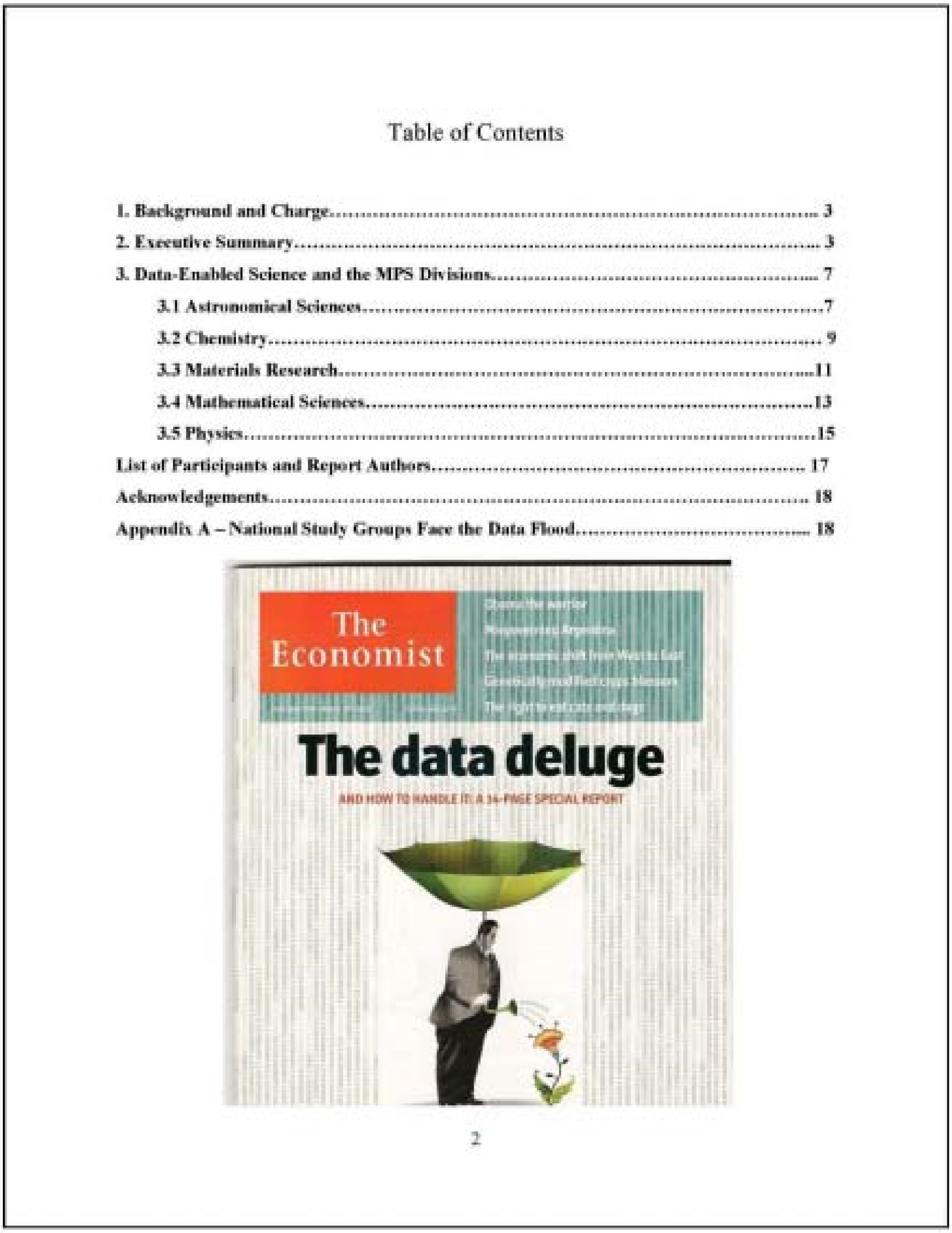
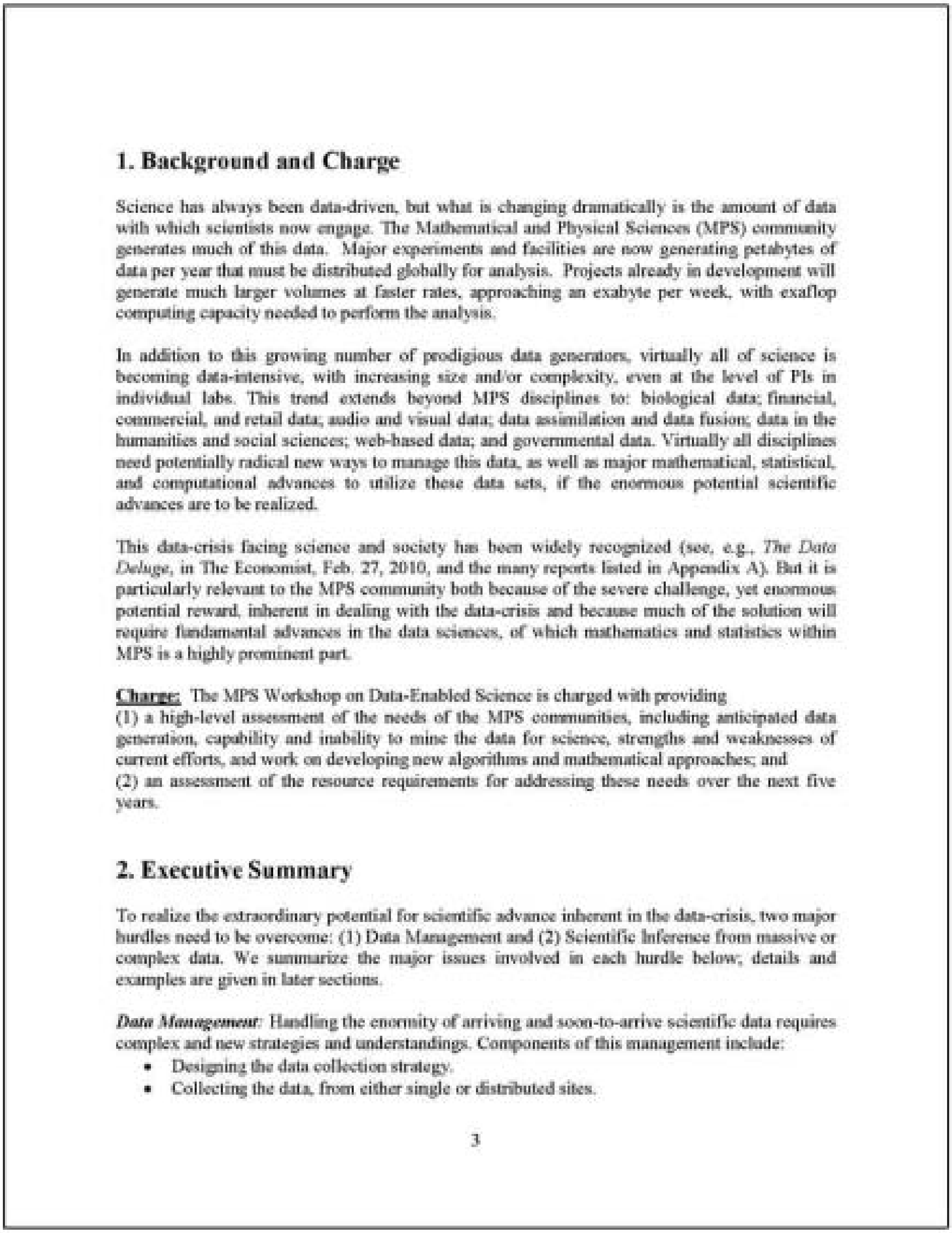


Table of Contents

2



1. Background and Charge

Science has always been data-driven, but what is changing dramatically is the amount of data with which scientists now engage. The Mathematical and Physical Sciences (MPS) community generates much of this data. Major experiments and facilities are now generating petabytes of data per year that must be distributed globally for analysis. Projects already in development will generate much larger volumes at faster rates, approaching an exabyte per week, with exaflop computing capacity needed to perform the analysis.

In addition to this growing number of prodigious data generators, virtually all of science is becoming data-intensive, with increasing size and/or complexity, even at the level of PIs in individual labs. This trend extends beyond MPS disciplines to: biological data; financial, commercial, and retail data; audio and visual data; data assimilation and data fusion; data in the humanities and social sciences; web-based data; and governmental data. Virtually all disciplines need potentially radical new ways to manage this data, as well as major mathematical, statistical, and computational advances to utilize these data sets, if the enormous potential scientific advances are to be realized.

This data-crisis facing science and society has been widely recognized (see, e.g., *The Data Deluge*, in The Economist, Feb. 27, 2010, and the many reports listed in Appendix A). But it is particularly relevant to the MPS community both because of the severe challenge, yet enormous potential reward, inherent in dealing with the data-crisis and because much of the solution will require fundamental advances in the data sciences, of which mathematics and statistics within MPS is a highly prominent part.

Charge: The MPS Workshop on Data-Enabled Science is charged with providing
(1) a high-level assessment of the needs of the MPS communities, including anticipated data generation, capability and inability to mine the data for science, strengths and weaknesses of current efforts, and work on developing new algorithms and mathematical approaches; and
(2) an assessment of the resource requirements for addressing these needs over the next five years.

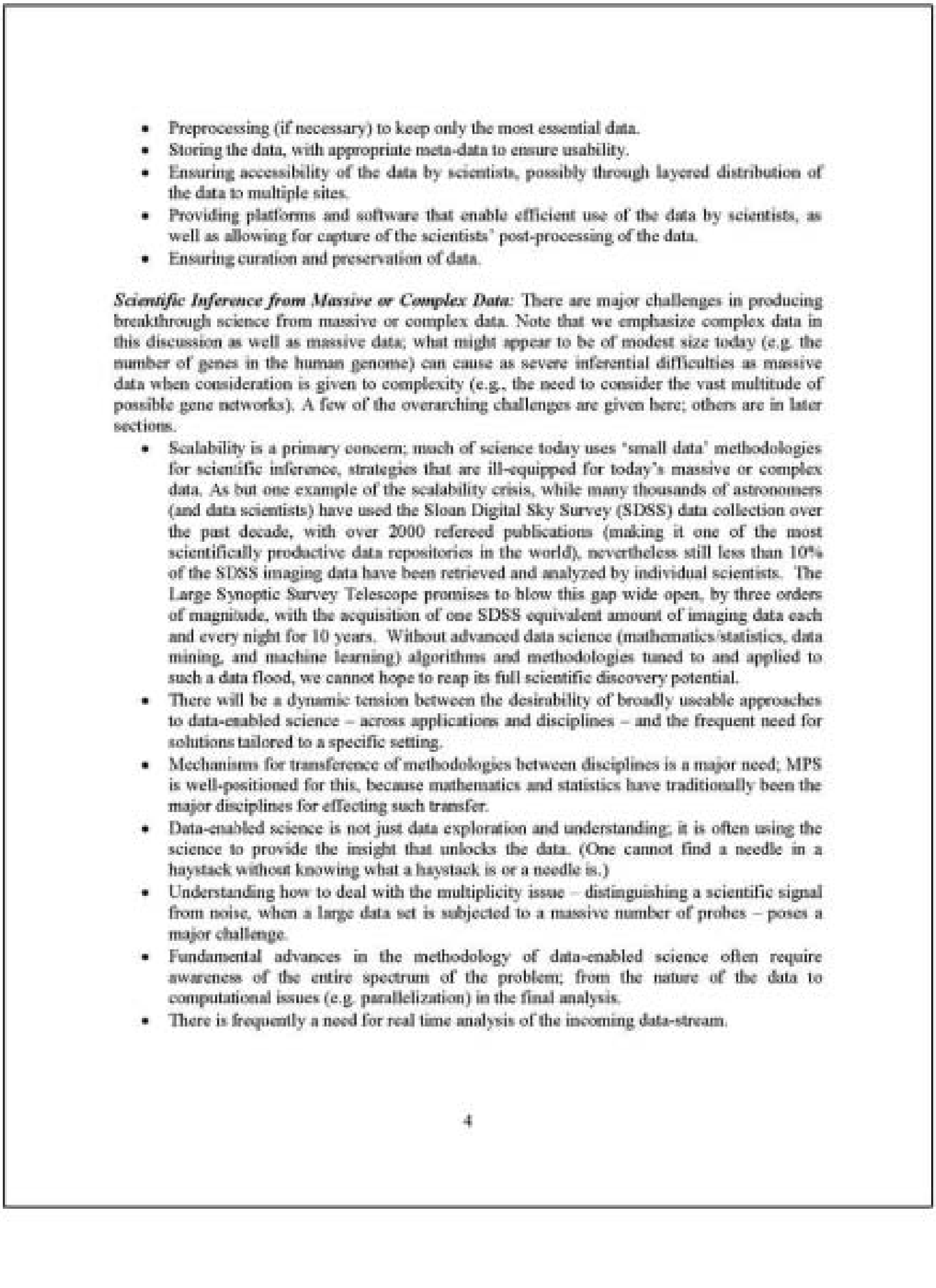
2. Executive Summary

To realize the extraordinary potential for scientific advance inherent in the data-crisis, two major hurdles need to be overcome: (1) Data Management and (2) Scientific Inference from massive or complex data. We summarize the major issues involved in each hurdle below; details and examples are given in later sections.

Data Management: Handling the enormity of arriving and soon-to-arrive scientific data requires complex and new strategies and understandings. Components of this management include:

- Designing the data collection strategy.
- Collecting the data, from either single or distributed sites.

3

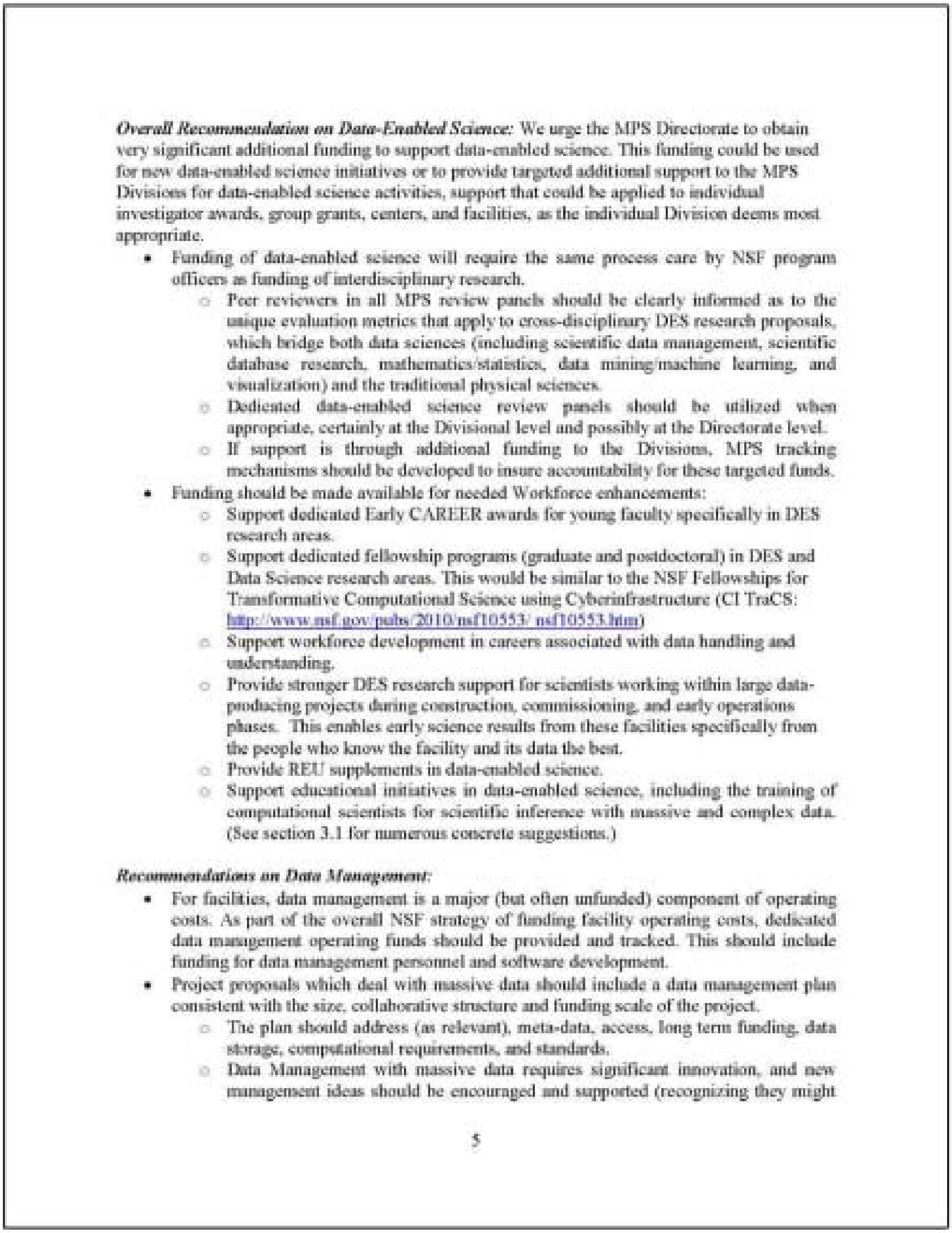


- Preprocessing (if necessary) to keep only the most essential data.
- Storing the data, with appropriate meta-data to ensure usability.
- Ensuring accessibility of the data by scientists, possibly through layered distribution of the data to multiple sites.
- Providing platforms and software that enable efficient use of the data by scientists, as well as allowing for capture of the scientists' post-processing of the data.
- Ensuring curation and preservation of data.

Scientific Inference from Massive or Complex Data: There are major challenges in producing breakthrough science from massive or complex data. Note that we emphasize complex data in this discussion as well as massive data; what might appear to be of modest size today (e.g. the number of genes in the human genome) can cause as severe inferential difficulties as massive data when consideration is given to complexity (e.g., the need to consider the vast multitude of possible gene networks). A few of the overarching challenges are given here; others are in later sections.

- Scalability is a primary concern; much of science today uses 'small data' methodologies for scientific inference, strategies that are ill-equipped for today's massive or complex data. As but one example of the scalability crisis, while many thousands of astronomers (and data scientists) have used the Sloan Digital Sky Survey (SDSS) data collection over the past decade, with over 2000 refereed publications (making it one of the most scientifically productive data repositories in the world), nevertheless still less than 10% of the SDSS imaging data have been retrieved and analyzed by individual scientists. The Large Synoptic Survey Telescope promises to blow this gap wide open, by three orders of magnitude, with the acquisition of one SDSS equivalent amount of imaging data each and every night for 10 years. Without advanced data science (mathematics/statistics, data mining, and machine learning) algorithms and methodologies tuned to and applied to such a data flood, we cannot hope to reap its full scientific discovery potential.
- There will be a dynamic tension between the desirability of broadly useable approaches to data-enabled science – across applications and disciplines – and the frequent need for solutions tailored to a specific setting.
- Mechanisms for transference of methodologies between disciplines is a major need; MPS is well-positioned for this, because mathematics and statistics have traditionally been the major disciplines for effecting such transfer.
- Data-enabled science is not just data exploration and understanding; it is often using the science to provide the insight that unlocks the data. (One cannot find a needle in a haystack without knowing what a haystack is or a needle is.)
- Understanding how to deal with the multiplicity issue – distinguishing a scientific signal from noise, when a large data set is subjected to a massive number of probes – poses a major challenge.
- Fundamental advances in the methodology of data-enabled science often require awareness of the entire spectrum of the problem; from the nature of the data to computational issues (e.g. parallelization) in the final analysis.
- There is frequently a need for real time analysis of the incoming data-stream.

4

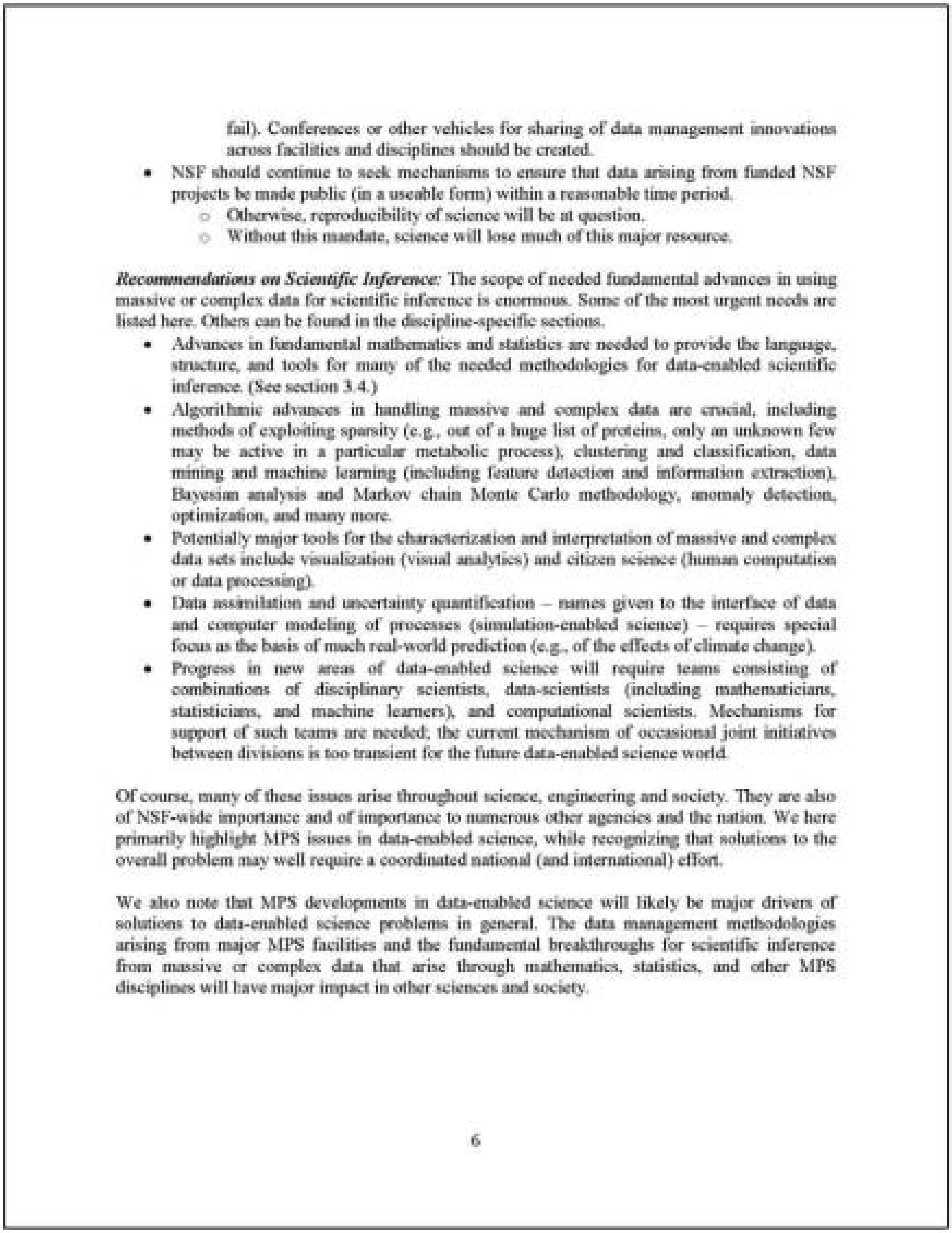


Overall Recommendation on Data-Enabled Science: We urge the MPS Directorate to obtain very significant additional funding to support data-enabled science. This funding could be used for new data-enabled science initiatives or to provide targeted additional support to the MPS Divisions for data-enabled science activities, support that could be applied to individual investigator awards, group grants, centers, and facilities, as the individual Division deems most appropriate.

- Funding of data-enabled science will require the same process care by NSF program officers as funding of interdisciplinary research.
  - Peer reviewers in all MPS review panels should be clearly informed as to the unique evaluation metrics that apply to cross-disciplinary DES research proposals, which bridge both data sciences (including scientific data management, scientific database research, mathematics/statistics, data mining/machine learning, and visualization) and the traditional physical sciences.
  - Dedicated data-enabled science review panels should be utilized when appropriate, certainly at the Divisional level and possibly at the Directorate level.
  - If support is through additional funding to the Divisions, MPS tracking mechanisms should be developed to insure accountability for these targeted funds.
- Funding should be made available for needed Workforce enhancements:
  - Support dedicated Early CAREER awards for young faculty specifically in DES research areas.
  - Support dedicated fellowship programs (graduate and postdoctoral) in DES and Data Science research areas. This would be similar to the NSF Fellowships for Transformative Computational Science using Cyberinfrastructure (CI TraCS: http://www.nsf.gov/pubs/2010/nsf10553/nsf10553.htm)
  - Support workforce development in careers associated with data handling and understanding.
  - Provide stronger DES research support for scientists working within large data-producing projects during construction, commissioning, and early operations phases. This enables early science results from these facilities specifically from the people who know the facility and its data the best.
  - Provide REU supplements in data-enabled science.
  - Support educational initiatives in data-enabled science, including the training of computational scientists for scientific inference with massive and complex data. (See section 3.1 for numerous concrete suggestions.)

Recommendations on Data Management:

- For facilities, data management is a major (but often unfunded) component of operating costs. As part of the overall NSF strategy of funding facility operating costs, dedicated data management operating funds should be provided and tracked. This should include funding for data management personnel and software development.
- Project proposals which deal with massive data should include a data management plan consistent with the size, collaborative structure and funding scale of the project.
  - The plan should address (as relevant), meta-data, access, long term funding, data storage, computational requirements, and standards.
  - Data Management with massive data requires significant innovation, and new management ideas should be encouraged and supported (recognizing they might



fail). Conferences or other vehicles for sharing of data management innovations across facilities and disciplines should be created.

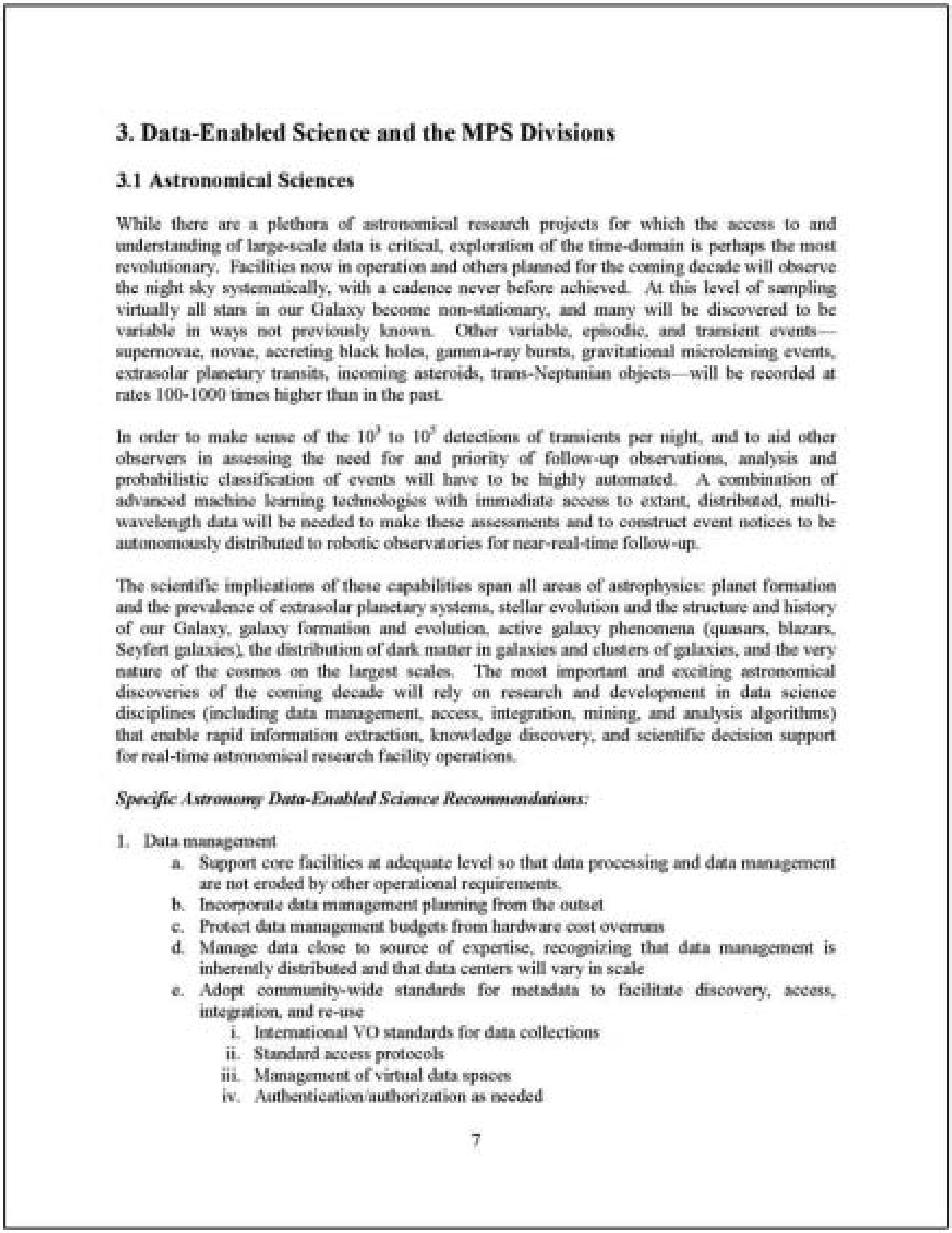
- NSF should continue to seek mechanisms to ensure that data arising from funded NSF projects be made public (in a useable form) within a reasonable time period.
    - Otherwise, reproducibility of science will be at question.
    - Without this mandate, science will lose much of this major resource.

Recommendations on Scientific Inference: The scope of needed fundamental advances in using massive or complex data for scientific inference is enormous. Some of the most urgent needs are listed here. Others can be found in the discipline-specific sections.

- Advances in fundamental mathematics and statistics are needed to provide the language, structure, and tools for many of the needed methodologies for data-enabled scientific inference. (See section 3.4.)
- Algorithmic advances in handling massive and complex data are crucial, including methods of exploiting sparsity (e.g., out of a huge list of proteins, only an unknown few may be active in a particular metabolic process), clustering and classification, data mining and machine learning (including feature detection and information extraction), Bayesian analysis and Markov chain Monte Carlo methodology, anomaly detection, optimization, and many more.
- Potentially major tools for the characterization and interpretation of massive and complex data sets include visualization (visual analytics) and citizen science (human computation or data processing).
- Data assimilation and uncertainty quantification – names given to the interface of data and computer modeling of processes (simulation-enabled science) – requires special focus as the basis of much real-world prediction (e.g., of the effects of climate change).
- Progress in new areas of data-enabled science will require teams consisting of combinations of disciplinary scientists, data-scientists (including mathematicians, statisticians, and machine learners), and computational scientists. Mechanisms for support of such teams are needed; the current mechanism of occasional joint initiatives between divisions is too transient for the future data-enabled science world.

Of course, many of these issues arise throughout science, engineering and society. They are also of NSF-wide importance and of importance to numerous other agencies and the nation. We here primarily highlight MPS issues in data-enabled science, while recognizing that solutions to the overall problem may well require a coordinated national (and international) effort.

We also note that MPS developments in data-enabled science will likely be major drivers of solutions to data-enabled science problems in general. The data management methodologies arising from major MPS facilities and the fundamental breakthroughs for scientific inference from massive or complex data that arise through mathematics, statistics, and other MPS disciplines will have major impact in other sciences and society.



3. Data-Enabled Science and the MPS Divisions

3.1 Astronomical Sciences

While there are a plethora of astronomical research projects for which the access to and understanding of large-scale data is critical, exploration of the time-domain is perhaps the most revolutionary. Facilities now in operation and others planned for the coming decade will observe the night sky systematically, with a cadence never before achieved. At this level of sampling virtually all stars in our Galaxy become non-stationary, and many will be discovered to be variable in ways not previously known. Other variable, episodic, and transient events—supernovae, novae, accreting black holes, gamma-ray bursts, gravitational microlensing events, extrasolar planetary transits, incoming asteroids, trans-Neptunian objects—will be recorded at rates 100-1000 times higher than in the past.

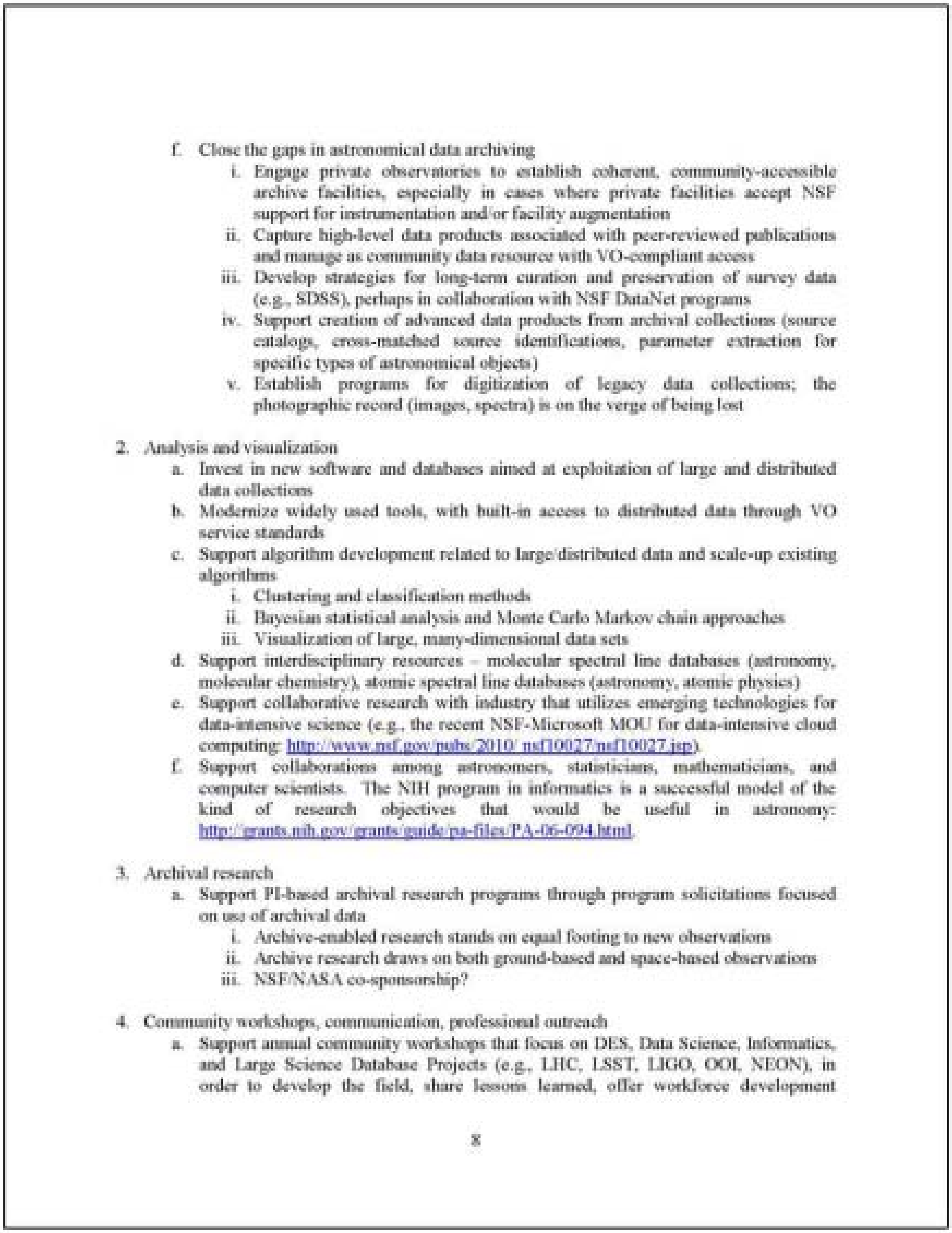
In order to make sense of the 10^5 to 10^7 detections of transients per night, and to aid other observers in assessing the need for and priority of follow-up observations, analysis and probabilistic classification of events will have to be highly automated. A combination of advanced machine learning technologies with immediate access to extant, distributed, multi-wavelength data will be needed to make these assessments and to construct event notices to be autonomously distributed to robotic observatories for near-real-time follow-up.

The scientific implications of these capabilities span all areas of astrophysics: planet formation and the prevalence of extrasolar planetary systems, stellar evolution and the structure and history of our Galaxy, galaxy formation and evolution, active galaxy phenomena (quasars, blazars, Seyfert galaxies), the distribution of dark matter in galaxies and clusters of galaxies, and the very nature of the cosmos on the largest scales. The most important and exciting astronomical discoveries of the coming decade will rely on research and development in data science disciplines (including data management, access, integration, mining, and analysis algorithms) that enable rapid information extraction, knowledge discovery, and scientific decision support for real-time astronomical research facility operations.

Specific Astronomy Data-Enabled Science Recommendations:

1. Data management
   a. Support core facilities at adequate level so that data processing and data management are not eroded by other operational requirements.
   b. Incorporate data management planning from the outset
   c. Protect data management budgets from hardware cost overruns
   d. Manage data close to source of expertise, recognizing that data management is inherently distributed and that data centers will vary in scale
   e. Adopt community-wide standards for metadata to facilitate discovery, access, integration, and re-use
      i. International VO standards for data collections
      ii. Standard access protocols
      iii. Management of virtual data spaces
      iv. Authentication/authorization as needed

7



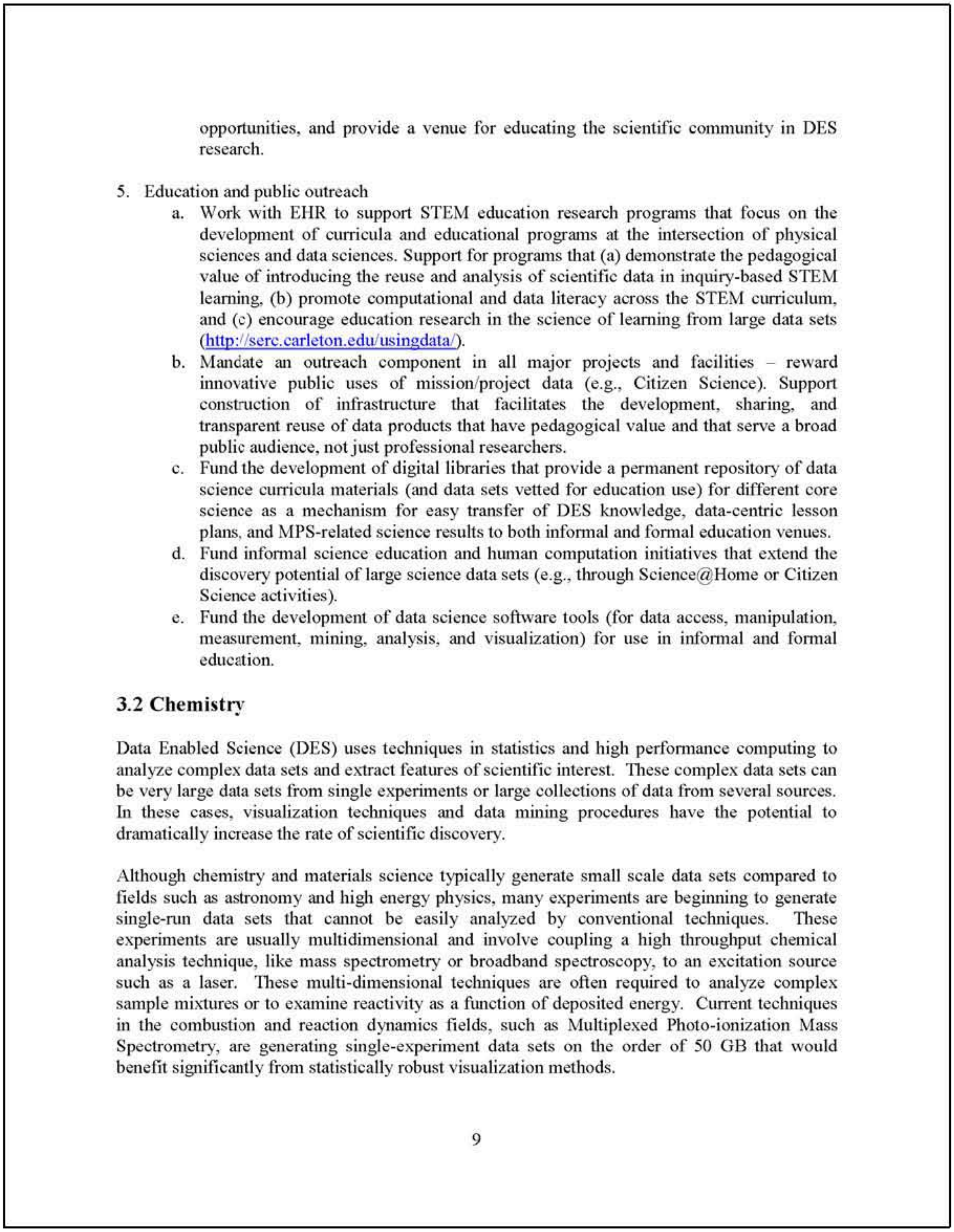
f. Close the gaps in astronomical data archiving
   i. Engage private observatories to establish coherent, community-accessible archive facilities, especially in cases where private facilities accept NSF support for instrumentation and/or facility augmentation
   ii. Capture high-level data products associated with peer-reviewed publications and manage as community data resource with VO-compliant access
   iii. Develop strategies for long-term curation and preservation of survey data (e.g., SDSS), perhaps in collaboration with NSF DataNet programs
   iv. Support creation of advanced data products from archival collections (source catalogs, cross-matched source identifications, parameter extraction for specific types of astronomical objects)
   v. Establish programs for digitization of legacy data collections; the photographic record (images, spectra) is on the verge of being lost

2. Analysis and visualization
   a. Invest in new software and databases aimed at exploitation of large and distributed data collections
   b. Modernize widely used tools, with built-in access to distributed data through VO service standards
   c. Support algorithm development related to large/distributed data and scale-up existing algorithms
      i. Clustering and classification methods
      ii. Bayesian statistical analysis and Monte Carlo Markov chain approaches
      iii. Visualization of large, many-dimensional data sets
   d. Support interdisciplinary resources – molecular spectral line databases (astronomy, molecular chemistry), atomic spectral line databases (astronomy, atomic physics)
   e. Support collaborative research with industry that utilizes emerging technologies for data-intensive science (e.g., the recent NSF-Microsoft MOU for data-intensive cloud computing: http://www.nsf.gov/pubs/2010/nsf10027/nsf10027.jsp).
   f. Support collaborations among astronomers, statisticians, mathematicians, and computer scientists. The NIH program in informatics is a successful model of the kind of research objectives that would be useful in astronomy: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html.

3. Archival research
   a. Support PI-based archival research programs through program solicitations focused on use of archival data
      i. Archive-enabled research stands on equal footing to new observations
      ii. Archive research draws on both ground-based and space-based observations
      iii. NSF/NASA co-sponsorship?

4. Community workshops, communication, professional outreach
   a. Support annual community workshops that focus on DES, Data Science, Informatics, and Large Science Database Projects (e.g., LHC, LSST, LIGO, OOI, NEON), in order to develop the field, share lessons learned, offer workforce development

8



opportunities, and provide a venue for educating the scientific community in DES research.

5. Education and public outreach
   a. Work with EHR to support STEM education research programs that focus on the development of curricula and educational programs at the intersection of physical sciences and data sciences. Support for programs that (a) demonstrate the pedagogical value of introducing the reuse and analysis of scientific data in inquiry-based STEM learning, (b) promote computational and data literacy across the STEM curriculum, and (c) encourage education research in the science of learning from large data sets (http://serc.carleton.edu/usingdata/).
   b. Mandate an outreach component in all major projects and facilities – reward innovative public uses of mission/project data (e.g., Citizen Science). Support construction of infrastructure that facilitates the development, sharing, and transparent reuse of data products that have pedagogical value and that serve a broad public audience, not just professional researchers.
   c. Fund the development of digital libraries that provide a permanent repository of data science curricula materials (and data sets vetted for education use) for different core science as a mechanism for easy transfer of DES knowledge, data-centric lesson plans, and MPS-related science results to both informal and formal education venues.
   d. Fund informal science education and human computation initiatives that extend the discovery potential of large science data sets (e.g., through Science@Home or Citizen Science activities).
   e. Fund the development of data science software tools (for data access, manipulation, measurement, mining, analysis, and visualization) for use in informal and formal education.

3.2 Chemistry

Data Enabled Science (DES) uses techniques in statistics and high performance computing to analyze complex data sets and extract features of scientific interest. These complex data sets can be very large data sets from single experiments or large collections of data from several sources. In these cases, visualization techniques and data mining procedures have the potential to dramatically increase the rate of scientific discovery.

Although chemistry and materials science typically generate small scale data sets compared to fields such as astronomy and high energy physics, many experiments are beginning to generate single-run data sets that cannot be easily analyzed by conventional techniques. These experiments are usually multidimensional and involve coupling a high throughput chemical analysis technique, like mass spectrometry or broadband spectroscopy, to an excitation source such as a laser. These multi-dimensional techniques are often required to analyze complex sample mixtures or to examine reactivity as a function of deposited energy. Current techniques in the combustion and reaction dynamics fields, such as Multiplexed Photo-ionization Mass Spectrometry, are generating single-experiment data sets on the order of 50 GB that would benefit significantly from statistically robust visualization methods.

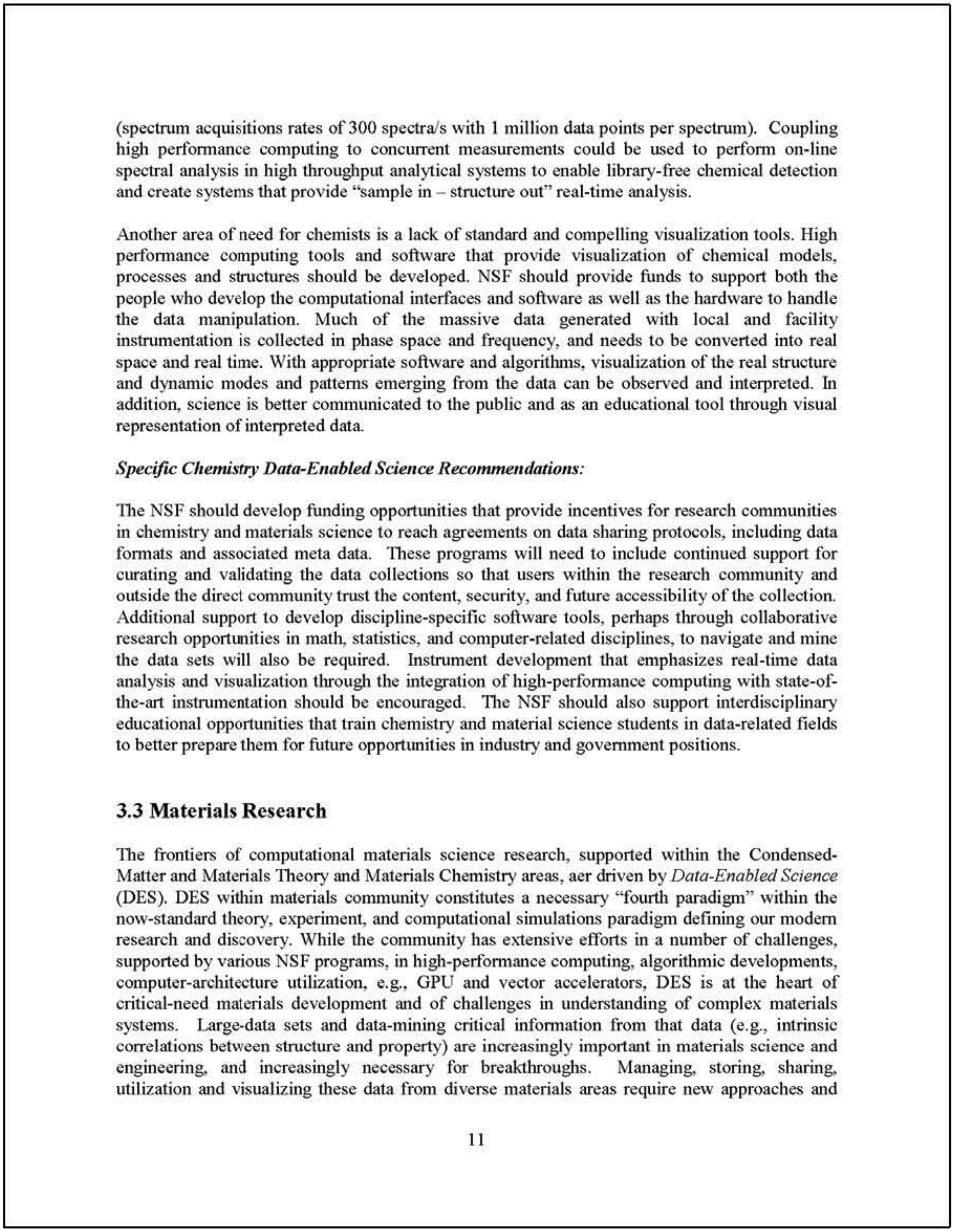
Several other areas of chemical and materials research are also producing large data sets that will continue to increase in size and complexity. In particular, molecular dynamics simulations in biochemistry and materials science generate large scale computational data from single laboratory studies. The use of graphics processing units in computational chemistry, for example, has led to simulations that produce terabytes of computational output per day. There are also large data sets in related fields of science, such as radio astronomy, that contain molecular information that require new analysis tools to extract the chemically useful information.

Finally, several of the industries that employ chemistry and material science Ph.D.'s are rapidly pursuing DES strategies to decrease product development cycles. Providing research experiences for graduate students will become increasingly important for preparing young scientists for the future workforce. Therefore, despite the "single laboratory" tradition of chemistry and material science research, issues in DES are already significant in chemistry and will continue to gain importance.

Special Needs for Chemistry and Material Science

As noted above, chemistry and materials science tend to perform research in a single-laboratory model. Increasingly, each individual laboratory is generating large scale data sets through either computational chemistry, large user facilities (such as SLAC, NIST or ORNL) or high throughput laboratory methods. However, the potentially greater opportunity for DES in these fields is the combination of research data from all groups in a research discipline. For example, a unified spectroscopic database from emerging high throughput spectroscopy methods based on frequency comb spectroscopy and direct digital spectroscopy could have a major impact on related fields of astronomy, environmental science, and analytical chemistry that rely on chemical identification by spectroscopy. Efforts are already underway in the computational chemistry community to create common data bases to permit reuse of these results (examples include iOpenShell (Krylov), the Structural Database (Johnson)). Unified collections of individual data sets in materials science and drug discovery could significantly increase the rate of discovery and add increased value to the individual laboratory data collections. The concept of unified data sets from whole communities of chemistry represents a major shift in the single-laboratory culture where data is often closely guarded.

A special area of DES with great potential in chemistry and material science fields is the combination of laboratory or facility measurements and computational chemistry to provide real-time chemical analysis. Many experiments in chemistry rely on theoretical analysis or computational simulation to interpret the experimental data. In almost all cases in chemistry, these tasks are performed separately and often by different research groups. For example, an experimental group will collect the data set and send it to a collaborator in computational chemistry for analysis. The possibility of closing this loop in real-time would make it possible to optimize experimental conditions in a single experimental run and, therefore, greatly decrease the time required to perform the crucial experiment to reveal the important chemistry. On-the-fly analysis methods are also needed to realize the full potential of new techniques like broadband spectroscopy using frequency combs or digital electronics. Spectrometers based on direct digital spectroscopy will soon be capable of measurement throughput of about 1 TB/hour



(spectrum acquisitions rates of 300 spectra/s with 1 million data points per spectrum). Coupling high performance computing to concurrent measurements could be used to perform on-line spectral analysis in high throughput analytical systems to enable library-free chemical detection and create systems that provide "sample in – structure out" real-time analysis.

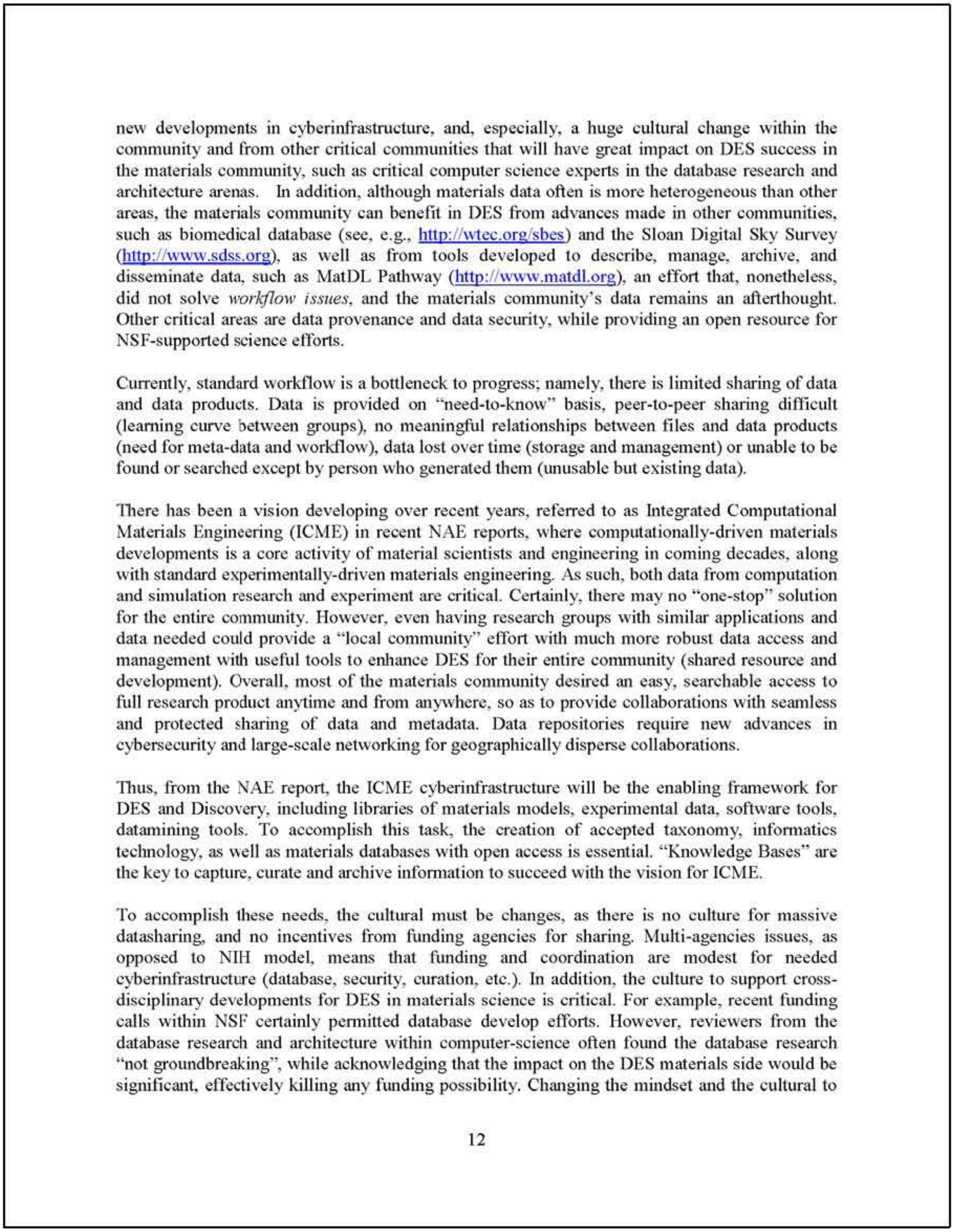
Another area of need for chemists is a lack of standard and compelling visualization tools. High performance computing tools and software that provide visualization of chemical models, processes and structures should be developed. NSF should provide funds to support both the people who develop the computational interfaces and software as well as the hardware to handle the data manipulation. Much of the massive data generated with local and facility instrumentation is collected in phase space and frequency, and needs to be converted into real space and real time. With appropriate software and algorithms, visualization of the real structure and dynamic modes and patterns emerging from the data can be observed and interpreted. In addition, science is better communicated to the public and as an educational tool through visual representation of interpreted data.

Specific Chemistry Data-Enabled Science Recommendations:

The NSF should develop funding opportunities that provide incentives for research communities in chemistry and materials science to reach agreements on data sharing protocols, including data formats and associated meta data. These programs will need to include continued support for curating and validating the data collections so that users within the research community and outside the direct community trust the content, security, and future accessibility of the collection. Additional support to develop discipline-specific software tools, perhaps through collaborative research opportunities in math, statistics, and computer-related disciplines, to navigate and mine the data sets will also be required. Instrument development that emphasizes real-time data analysis and visualization through the integration of high-performance computing with state-of-the-art instrumentation should be encouraged. The NSF should also support interdisciplinary educational opportunities that train chemistry and material science students in data-related fields to better prepare them for future opportunities in industry and government positions.

3.3 Materials Research

The frontiers of computational materials science research, supported within the Condensed-Matter and Materials Theory and Materials Chemistry areas, aer driven by *Data-Enabled Science* (DES). DES within materials community constitutes a necessary "fourth paradigm" within the now-standard theory, experiment, and computational simulations paradigm defining our modern research and discovery. While the community has extensive efforts in a number of challenges, supported by various NSF programs, in high-performance computing, algorithmic developments, computer-architecture utilization, e.g., GPU and vector accelerators, DES is at the heart of critical-need materials development and of challenges in understanding of complex materials systems. Large-data sets and data-mining critical information from that data (e.g., intrinsic correlations between structure and property) are increasingly important in materials science and engineering, and increasingly necessary for breakthroughs. Managing, storing, sharing, utilization and visualizing these data from diverse materials areas require new approaches and



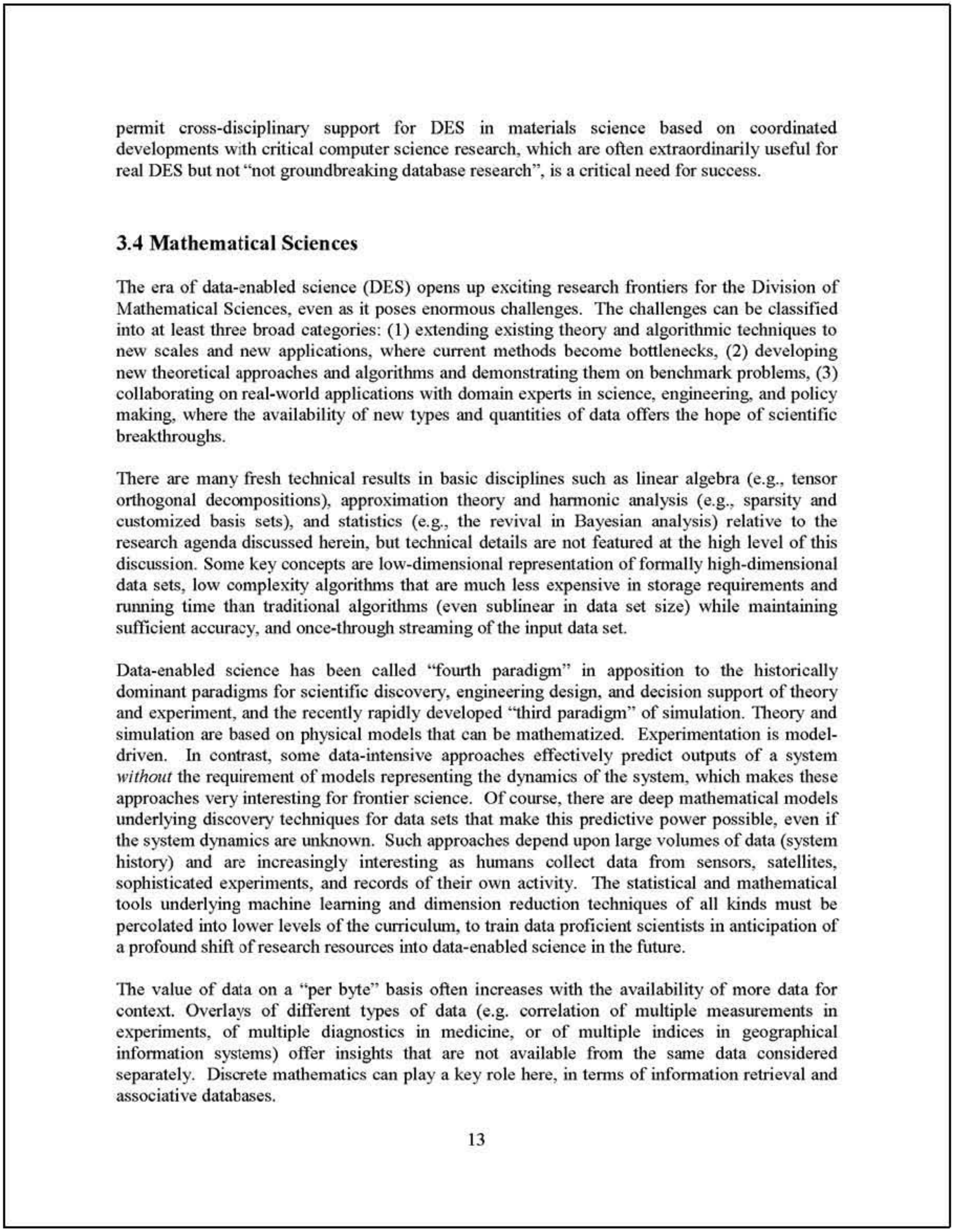
new developments in cyberinfrastructure, and, especially, a huge cultural change within the community and from other critical communities that will have great impact on DES success in the materials community, such as critical computer science experts in the database research and architecture arenas. In addition, although materials data often is more heterogeneous than other areas, the materials community can benefit in DES from advances made in other communities, such as biomedical database (see, e.g., http://wtec.org/sbes) and the Sloan Digital Sky Survey (http://www.sdss.org), as well as from tools developed to describe, manage, archive, and disseminate data, such as MatDL Pathway (http://www.matdl.org), an effort that, nonetheless, did not solve *workflow issues*, and the materials community's data remains an afterthought. Other critical areas are data provenance and data security, while providing an open resource for NSF-supported science efforts.

Currently, standard workflow is a bottleneck to progress; namely, there is limited sharing of data and data products. Data is provided on "need-to-know" basis, peer-to-peer sharing difficult (learning curve between groups), no meaningful relationships between files and data products (need for meta-data and workflow), data lost over time (storage and management) or unable to be found or searched except by person who generated them (unusable but existing data).

There has been a vision developing over recent years, referred to as Integrated Computational Materials Engineering (ICME) in recent NAE reports, where computationally-driven materials developments is a core activity of material scientists and engineering in coming decades, along with standard experimentally-driven materials engineering. As such, both data from computation and simulation research and experiment are critical. Certainly, there may no "one-stop" solution for the entire community. However, even having research groups with similar applications and data needed could provide a "local community" effort with much more robust data access and management with useful tools to enhance DES for their entire community (shared resource and development). Overall, most of the materials community desired an easy, searchable access to full research product anytime and from anywhere, so as to provide collaborations with seamless and protected sharing of data and metadata. Data repositories require new advances in cybersecurity and large-scale networking for geographically disperse collaborations.

Thus, from the NAE report, the ICME cyberinfrastructure will be the enabling framework for DES and Discovery, including libraries of materials models, experimental data, software tools, datamining tools. To accomplish this task, the creation of accepted taxonomy, informatics technology, as well as materials databases with open access is essential. "Knowledge Bases" are the key to capture, curate and archive information to succeed with the vision for ICME.

To accomplish these needs, the cultural must be changes, as there is no culture for massive datasharing, and no incentives from funding agencies for sharing. Multi-agencies issues, as opposed to NIH model, means that funding and coordination are modest for needed cyberinfrastructure (database, security, curation, etc.). In addition, the culture to support cross-disciplinary developments for DES in materials science is critical. For example, recent funding calls within NSF certainly permitted database develop efforts. However, reviewers from the database research and architecture within computer-science often found the database research "not groundbreaking", while acknowledging that the impact on the DES materials side would be significant, effectively killing any funding possibility. Changing the mindset and the cultural to



permit cross-disciplinary support for DES in materials science based on coordinated developments with critical computer science research, which are often extraordinarily useful for real DES but not "not groundbreaking database research", is a critical need for success.

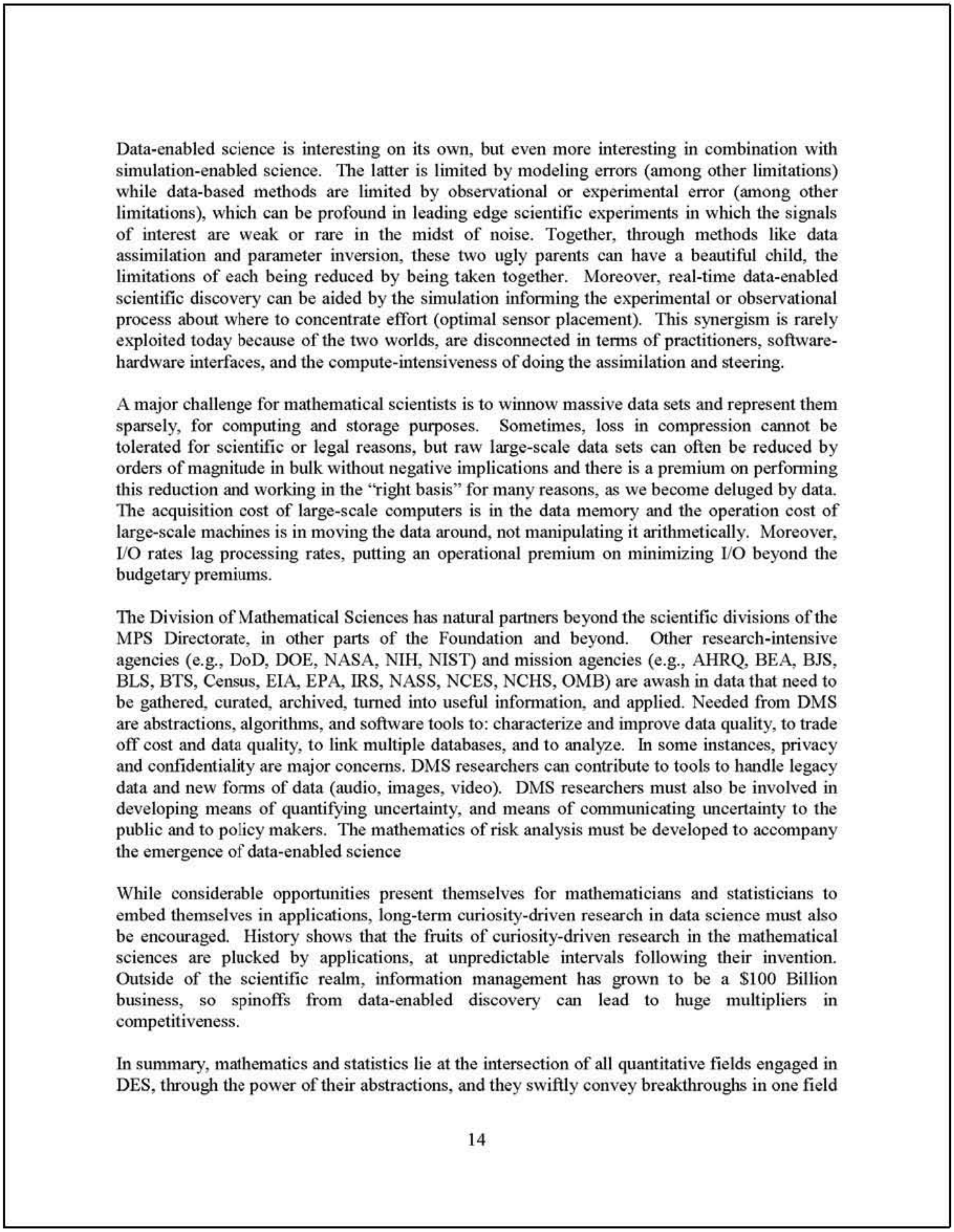
3.4 Mathematical Sciences

The era of data-enabled science (DES) opens up exciting research frontiers for the Division of Mathematical Sciences, even as it poses enormous challenges. The challenges can be classified into at least three broad categories: (1) extending existing theory and algorithmic techniques to new scales and new applications, where current methods become bottlenecks, (2) developing new theoretical approaches and algorithms and demonstrating them on benchmark problems, (3) collaborating on real-world applications with domain experts in science, engineering, and policy making, where the availability of new types and quantities of data offers the hope of scientific breakthroughs.

There are many fresh technical results in basic disciplines such as linear algebra (e.g., tensor orthogonal decompositions), approximation theory and harmonic analysis (e.g., sparsity and customized basis sets), and statistics (e.g., the revival in Bayesian analysis) relative to the research agenda discussed herein, but technical details are not featured at the high level of this discussion. Some key concepts are low-dimensional representation of formally high-dimensional data sets, low complexity algorithms that are much less expensive in storage requirements and running time than traditional algorithms (even sublinear in data set size) while maintaining sufficient accuracy, and once-through streaming of the input data set.

Data-enabled science has been called "fourth paradigm" in apposition to the historically dominant paradigms for scientific discovery, engineering design, and decision support of theory and experiment, and the recently rapidly developed "third paradigm" of simulation. Theory and simulation are based on physical models that can be mathematized. Experimentation is model-driven. In contrast, some data-intensive approaches effectively predict outputs of a system *without* the requirement of models representing the dynamics of the system, which makes these approaches very interesting for frontier science. Of course, there are deep mathematical models underlying discovery techniques for data sets that make this predictive power possible, even if the system dynamics are unknown. Such approaches depend upon large volumes of data (system history) and are increasingly interesting as humans collect data from sensors, satellites, sophisticated experiments, and records of their own activity. The statistical and mathematical tools underlying machine learning and dimension reduction techniques of all kinds must be percolated into lower levels of the curriculum, to train data proficient scientists in anticipation of a profound shift of research resources into data-enabled science in the future.

The value of data on a "per byte" basis often increases with the availability of more data for context. Overlays of different types of data (e.g. correlation of multiple measurements in experiments, of multiple diagnostics in medicine, or of multiple indices in geographical information systems) offer insights that are not available from the same data considered separately. Discrete mathematics can play a key role here, in terms of information retrieval and associative databases.



Data-enabled science is interesting on its own, but even more interesting in combination with simulation-enabled science. The latter is limited by modeling errors (among other limitations) while data-based methods are limited by observational or experimental error (among other limitations), which can be profound in leading edge scientific experiments in which the signals of interest are weak or rare in the midst of noise. Together, through methods like data assimilation and parameter inversion, these two ugly parents can have a beautiful child, the limitations of each being reduced by being taken together. Moreover, real-time data-enabled scientific discovery can be aided by the simulation informing the experimental or observational process about where to concentrate effort (optimal sensor placement). This synergism is rarely exploited today because of the two worlds, are disconnected in terms of practitioners, software-hardware interfaces, and the compute-intensiveness of doing the assimilation and steering.

A major challenge for mathematical scientists is to winnow massive data sets and represent them sparsely, for computing and storage purposes. Sometimes, loss in compression cannot be tolerated for scientific or legal reasons, but raw large-scale data sets can often be reduced by orders of magnitude in bulk without negative implications and there is a premium on performing this reduction and working in the "right basis" for many reasons, as we become deluged by data. The acquisition cost of large-scale computers is in the data memory and the operation cost of large-scale machines is in moving the data around, not manipulating it arithmetically. Moreover, I/O rates lag processing rates, putting an operational premium on minimizing I/O beyond the budgetary premiums.

The Division of Mathematical Sciences has natural partners beyond the scientific divisions of the MPS Directorate, in other parts of the Foundation and beyond. Other research-intensive agencies (e.g., DoD, DOE, NASA, NIH, NIST) and mission agencies (e.g., AHRQ, BEA, BJS, BLS, BTS, Census, EIA, EPA, IRS, NASS, NCES, NCHS, OMB) are awash in data that need to be gathered, curated, archived, turned into useful information, and applied. Needed from DMS are abstractions, algorithms, and software tools to: characterize and improve data quality, to trade off cost and data quality, to link multiple databases, and to analyze. In some instances, privacy and confidentiality are major concerns. DMS researchers can contribute to tools to handle legacy data and new forms of data (audio, images, video). DMS researchers must also be involved in developing means of quantifying uncertainty, and means of communicating uncertainty to the public and to policy makers. The mathematics of risk analysis must be developed to accompany the emergence of data-enabled science

While considerable opportunities present themselves for mathematicians and statisticians to embed themselves in applications, long-term curiosity-driven research in data science must also be encouraged. History shows that the fruits of curiosity-driven research in the mathematical sciences are plucked by applications, at unpredictable intervals following their invention. Outside of the scientific realm, information management has grown to be a $100 Billion business, so spinoffs from data-enabled discovery can lead to huge multipliers in competitiveness.

In summary, mathematics and statistics lie at the intersection of all quantitative fields engaged in DES, through the power of their abstractions, and they swiftly convey breakthroughs in one field into related ones. Individuals in DMS are often involved with the frontiers of DES both internally to the discipline and in interdisciplinary contexts. Growing numbers in the mathematical sciences community wish to be involved in DES problems, which has led to some of the most innovative, prize-winning developments in mathematics and statistics in recent years and some of the greatest fun. Impediments to be addressed by MPSAC could include the difficulty of securing postdoctoral funding (especially in statistics) and limited opportunities for interdisciplinary engagements as co-PIs on project proposals to NSF, since projects that are truly collaborative may present particular challenges to review panels.

3.5 Physics

Large data sets are a familiar component of physics research. In recent years, LIGO has acquired about two petabytes of data. With the Large Hadron Collider (LHC) reaching interesting beam energies, particle physics is preparing for the impending data tsunami which will generate about 700 MB of data per second. And this does not include simulated data, which could easily double or triple the data rate.

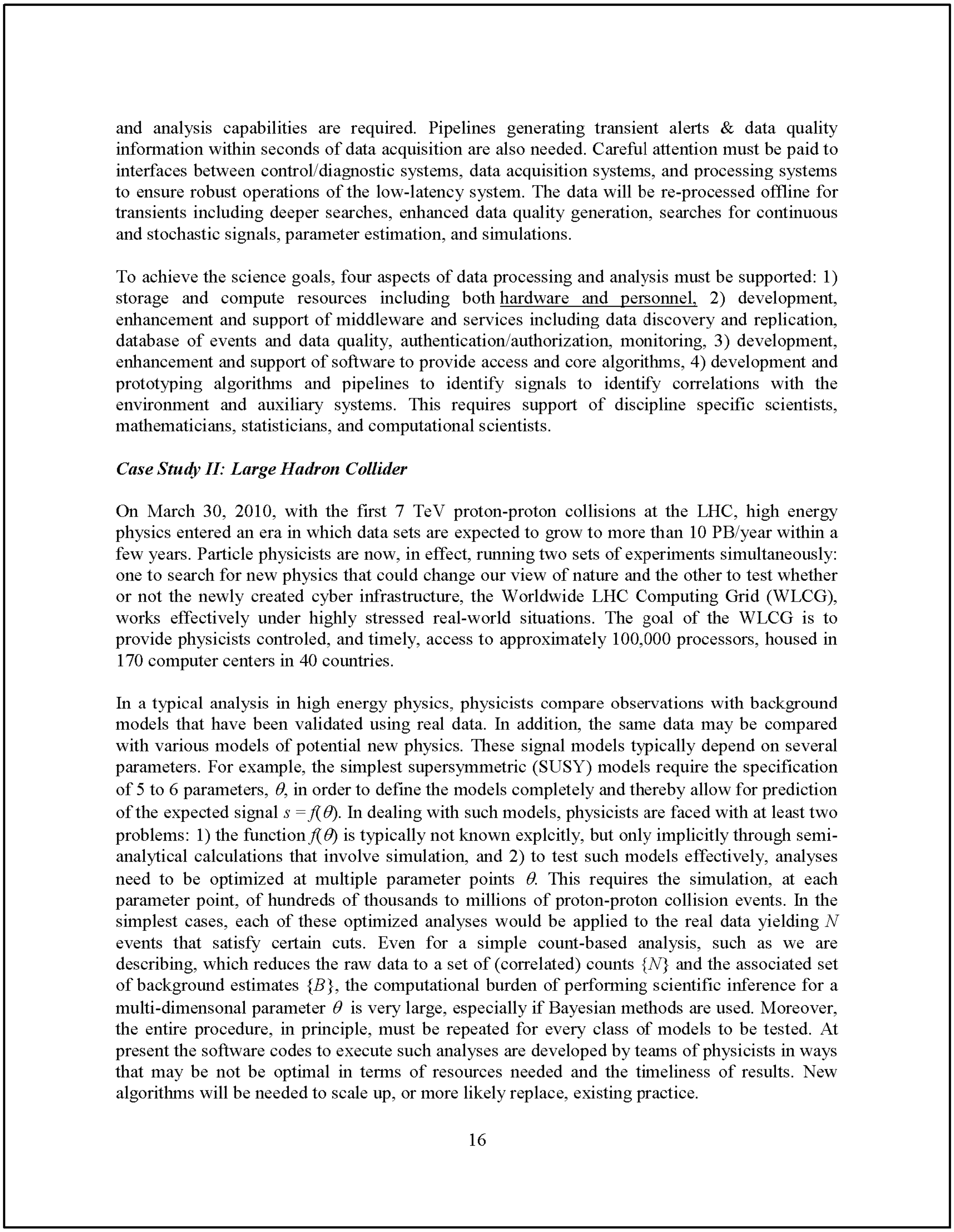
These big experiments are not the only data-enabled physics, however. The ability to simulate complex physical systems is also advancing rapidly. The output of these simulations will grow in size and complexity as more physics is included in the simulations. Moreover, single investigator experimental programs can easily acquire large amounts of data and many would benefit from better algorithm, software, and even data sharing formats.

The scientific pay-off of these data-intensive projects is bounded by the ability to process and analyze the data at the rate they are acquired.

Case Study I: Gravitational-wave Astronomy (LIGO)

The scientific pay-off of LIGO is bounded by the ability to process and analyze the data at the rate they are acquired. Over the past decade, LIGO has acquired 2 petabytes of data. The scientific collaboration adopted an hierarchical grid model for data storage and computation in which raw data is archived in Tier-0 data centers and centrally aggregated to a Tier 1 from which reduced data is moved to Tier-2 (regional compute centers) and Tier-3 (university compute centers). A similar structure has been adopted by the LHC experiments.

Over the next five years, the Advanced LIGO instrumentats (aLIGO) will be installed and begin operating. LIGO has partnered with Virgo, a French-Italian gravitational-wave detector project, and with GEO, a British-German detector project, to form a global network of gravitational-wave detectors. The goals of aLIGO are to test relativistic gravity and to develop gravitational-wave detection as an astronomical probe. aLIGO operations will span the transition from rare detections to routine astronomical observations. In stable operations, aLIGO will generate about 1 PB of raw data per year which needs to be replicated between the geographically distributed observatories and the compute centers at the same rate as it is acquired. A number of processed data products are planned including reduced data sets for scientific analysis, event databases, and astronomical alerts when transient events are identified. Robust online and offline data handling



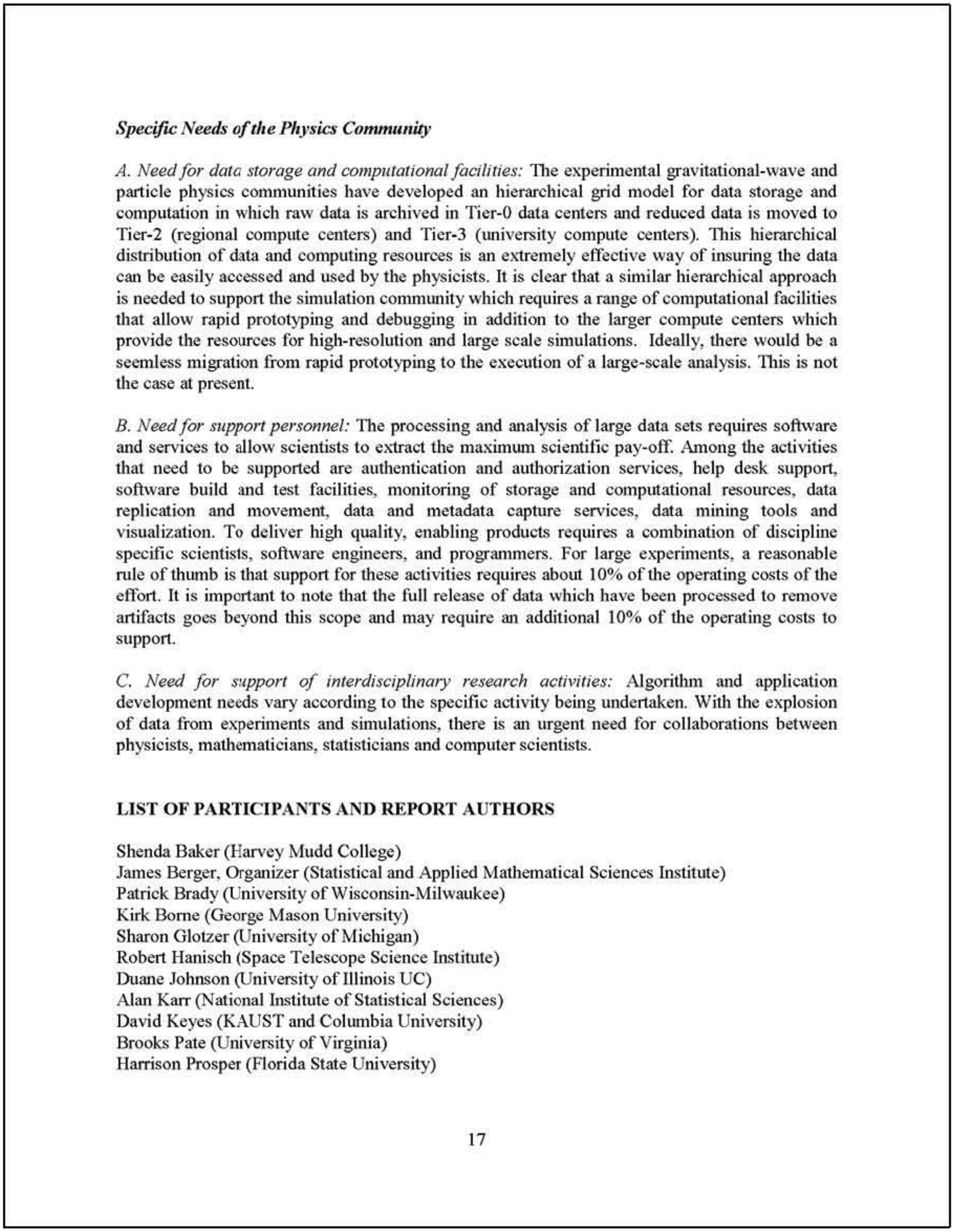
and analysis capabilities are required. Pipelines generating transient alerts & data quality information within seconds of data acquisition are also needed. Careful attention must be paid to interfaces between control/diagnostic systems, data acquisition systems, and processing systems to ensure robust operations of the low-latency system. The data will be re-processed offline for transients including deeper searches, enhanced data quality generation, searches for continuous and stochastic signals, parameter estimation, and simulations.

To achieve the science goals, four aspects of data processing and analysis must be supported: 1) storage and compute resources including both hardware and personnel, 2) development, enhancement and support of middleware and services including data discovery and replication, database of events and data quality, authentication/authorization, monitoring, 3) development, enhancement and support of software to provide access and core algorithms, 4) development and prototyping algorithms and pipelines to identify signals to identify correlations with the environment and auxiliary systems. This requires support of discipline specific scientists, mathematicians, statisticians, and computational scientists.

Case Study II: Large Hadron Collider

On March 30, 2010, with the first 7 TeV proton-proton collisions at the LHC, high energy physics entered an era in which data sets are expected to grow to more than 10 PB/year within a few years. Particle physicists are now, in effect, running two sets of experiments simultaneously: one to search for new physics that could change our view of nature and the other to test whether or not the newly created cyber infrastructure, the Worldwide LHC Computing Grid (WLCG), works effectively under highly stressed real-world situations. The goal of the WLCG is to provide physicists controled, and timely, access to approximately 100,000 processors, housed in 170 computer centers in 40 countries.

In a typical analysis in high energy physics, physicists compare observations with background models that have been validated using real data. In addition, the same data may be compared with various models of potential new physics. These signal models typically depend on several parameters. For example, the simplest supersymmetric (SUSY) models require the specification of 5 to 6 parameters, θ, in order to define the models completely and thereby allow for prediction of the expected signal $s = f(\theta)$. In dealing with such models, physicists are faced with at least two problems: 1) the function $f(\theta)$ is typically not known explcitly, but only implicitly through semi-analytical calculations that involve simulation, and 2) to test such models effectively, analyses need to be optimized at multiple parameter points θ. This requires the simulation, at each parameter point, of hundreds of thousands to millions of proton-proton collision events. In the simplest cases, each of these optimized analyses would be applied to the real data yielding N events that satisfy certain cuts. Even for a simple count-based analysis, such as we are describing, which reduces the raw data to a set of (correlated) counts $\{N\}$ and the associated set of background estimates $\{B\}$, the computational burden of performing scientific inference for a multi-dimensonal parameter θ is very large, especially if Bayesian methods are used. Moreover, the entire procedure, in principle, must be repeated for every class of models to be tested. At present the software codes to execute such analyses are developed by teams of physicists in ways that may be not be optimal in terms of resources needed and the timeliness of results. New algorithms will be needed to scale up, or more likely replace, existing practice.



Specific Needs of the Physics Community

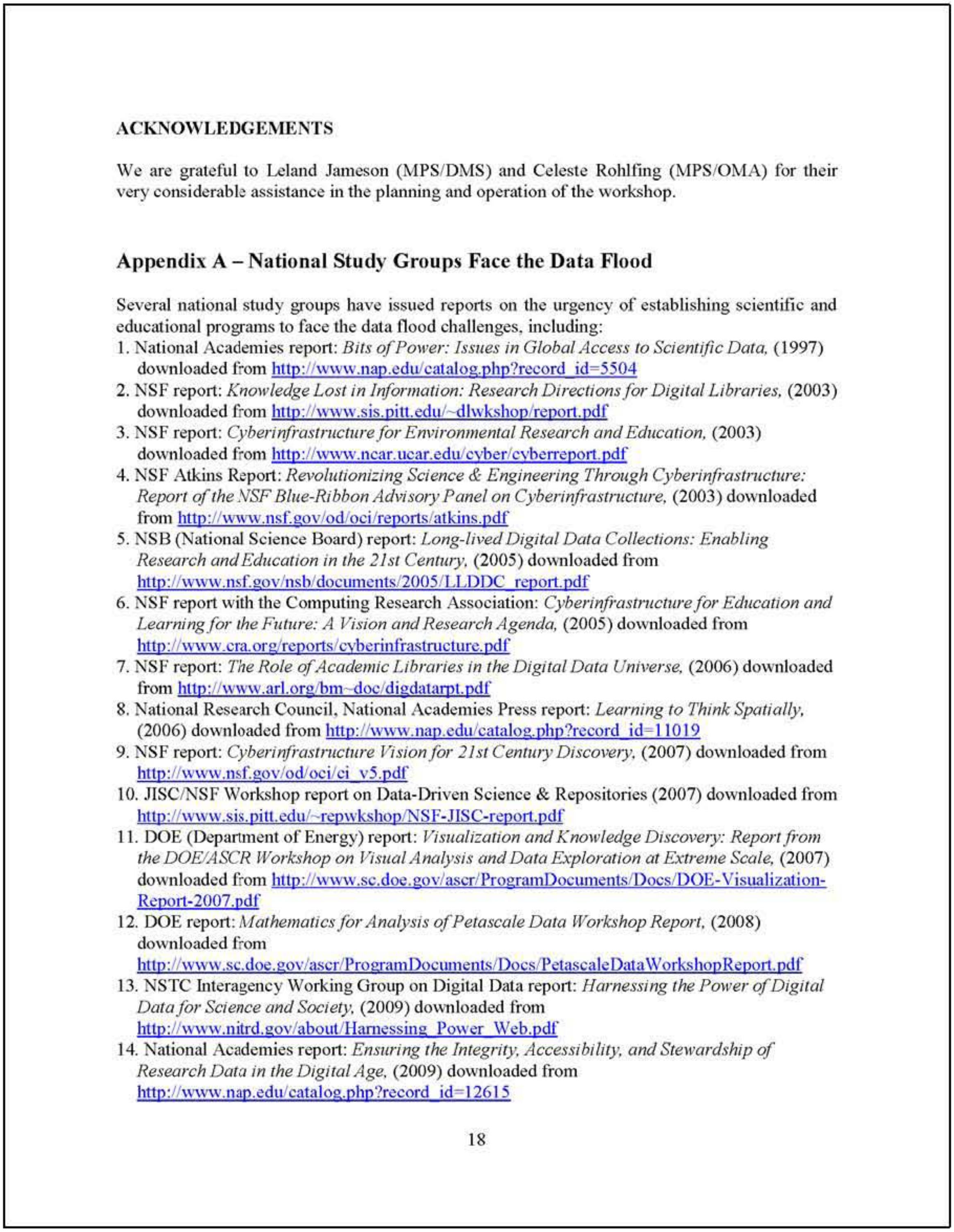
A. Need for data storage and computational facilities: The experimental gravitational-wave and particle physics communities have developed an hierarchical grid model for data storage and computation in which raw data is archived in Tier-0 data centers and reduced data is moved to Tier-2 (regional compute centers) and Tier-3 (university compute centers). This hierarchical distribution of data and computing resources is an extremely effective way of insuring the data can be easily accessed and used by the physicists. It is clear that a similar hierarchical approach is needed to support the simulation community which requires a range of computational facilities that allow rapid prototyping and debugging in addition to the larger compute centers which provide the resources for high-resolution and large scale simulations. Ideally, there would be a seemless migration from rapid prototyping to the execution of a large-scale analysis. This is not the case at present.

B. Need for support personnel: The processing and analysis of large data sets requires software and services to allow scientists to extract the maximum scientific pay-off. Among the activities that need to be supported are authentication and authorization services, help desk support, software build and test facilities, monitoring of storage and computational resources, data replication and movement, data and metadata capture services, data mining tools and visualization. To deliver high quality, enabling products requires a combination of discipline specific scientists, software engineers, and programmers. For large experiments, a reasonable rule of thumb is that support for these activities requires about 10% of the operating costs of the effort. It is important to note that the full release of data which have been processed to remove artifacts goes beyond this scope and may require an additional 10% of the operating costs to support.

C. Need for support of interdisciplinary research activities: Algorithm and application development needs vary according to the specific activity being undertaken. With the explosion of data from experiments and simulations, there is an urgent need for collaborations between physicists, mathematicians, statisticians and computer scientists.

LIST OF PARTICIPANTS AND REPORT AUTHORS

Shenda Baker (Harvey Mudd College)
James Berger, Organizer (Statistical and Applied Mathematical Sciences Institute)
Patrick Brady (University of Wisconsin-Milwaukee)
Kirk Borne (George Mason University)
Sharon Glotzer (University of Michigan)
Robert Hanisch (Space Telescope Science Institute)
Duane Johnson (University of Illinois UC)
Alan Karr (National Institute of Statistical Sciences)
David Keyes (KAUST and Columbia University)
Brooks Pate (University of Virginia)
Harrison Prosper (Florida State University)



ACKNOWLEDGEMENTS

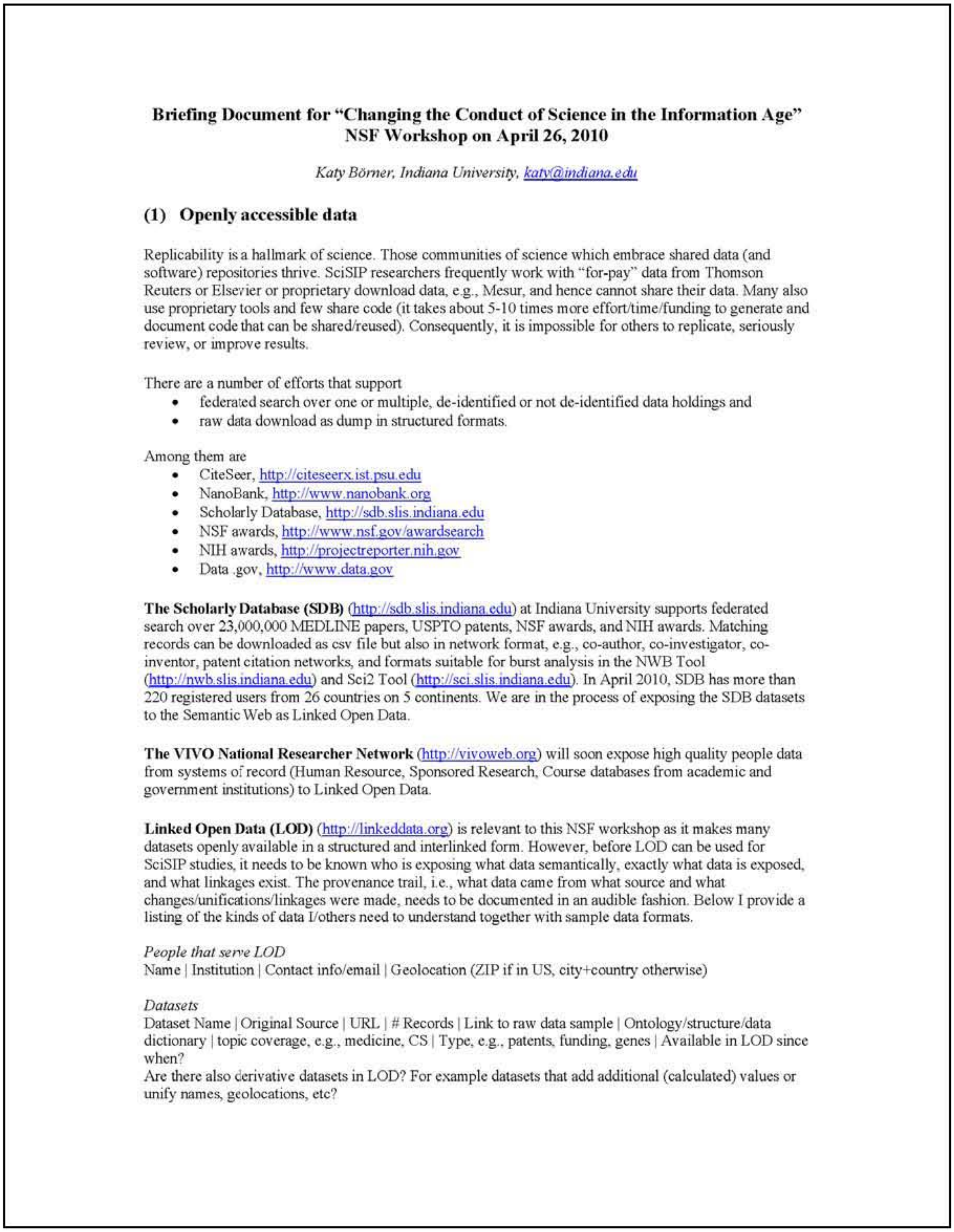
We are grateful to Leland Jameson (MPS/DMS) and Celeste Rohlfing (MPS/OMA) for their very considerable assistance in the planning and operation of the workshop.

Appendix A – National Study Groups Face the Data Flood

Several national study groups have issued reports on the urgency of establishing scientific and educational programs to face the data flood challenges, including:

1. National Academies report: *Bits of Power: Issues in Global Access to Scientific Data,* (1997) downloaded from http://www.nap.edu/catalog.php?record_id=5504
2. NSF report: *Knowledge Lost in Information: Research Directions for Digital Libraries,* (2003) downloaded from http://www.sis.pitt.edu/~dlwkshop/report.pdf
3. NSF report: *Cyberinfrastructure for Environmental Research and Education,* (2003) downloaded from http://www.ncar.ucar.edu/cyber/cyberreport.pdf
4. NSF Atkins Report: *Revolutionizing Science & Engineering Through Cyberinfrastructure: Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure,* (2003) downloaded from http://www.nsf.gov/od/oci/reports/atkins.pdf
5. NSB (National Science Board) report: *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century,* (2005) downloaded from http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf
6. NSF report with the Computing Research Association: *Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda,* (2005) downloaded from http://www.cra.org/reports/cyberinfrastructure.pdf
7. NSF report: *The Role of Academic Libraries in the Digital Data Universe,* (2006) downloaded from http://www.arl.org/bm~doc/digdatarpt.pdf
8. National Research Council, National Academies Press report: *Learning to Think Spatially,* (2006) downloaded from http://www.nap.edu/catalog.php?record_id=11019
9. NSF report: *Cyberinfrastructure Vision for 21st Century Discovery,* (2007) downloaded from http://www.nsf.gov/od/oci/ci_v5.pdf
10. JISC/NSF Workshop report on Data-Driven Science & Repositories (2007) downloaded from http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf
11. DOE (Department of Energy) report: *Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale,* (2007) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/DOE-Visualization-Report-2007.pdf
12. DOE report: *Mathematics for Analysis of Petascale Data Workshop Report,* (2008) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf
13. NSTC Interagency Working Group on Digital Data report: *Harnessing the Power of Digital Data for Science and Society,* (2009) downloaded from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf
14. National Academies report: *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age,* (2009) downloaded from http://www.nap.edu/catalog.php?record_id=12615

Börner, Katy, "Briefing Document for Changing the Conduct of Science in the Information Age"



Briefing Document for "Changing the Conduct of Science in the Information Age"
NSF Workshop on April 26, 2010

Katy Börner, Indiana University, katy@indiana.edu

(1) Openly accessible data

Replicability is a hallmark of science. Those communities of science which embrace shared data (and software) repositories thrive. SciSIP researchers frequently work with "for-pay" data from Thomson Reuters or Elsevier or proprietary download data, e.g., Mesur, and hence cannot share their data. Many also use proprietary tools and few share code (it takes about 5-10 times more effort/time/funding to generate and document code that can be shared/reused). Consequently, it is impossible for others to replicate, seriously review, or improve results.

There are a number of efforts that support

- federated search over one or multiple, de-identified or not de-identified data holdings and
- raw data download as dump in structured formats.

Among them are

- CiteSeer, http://citeseerx.ist.psu.edu
- NanoBank, http://www.nanobank.org
- Scholarly Database, http://sdb.slis.indiana.edu
- NSF awards, http://www.nsf.gov/awardsearch
- NIH awards, http://projectreporter.nih.gov
- Data .gov, http://www.data.gov

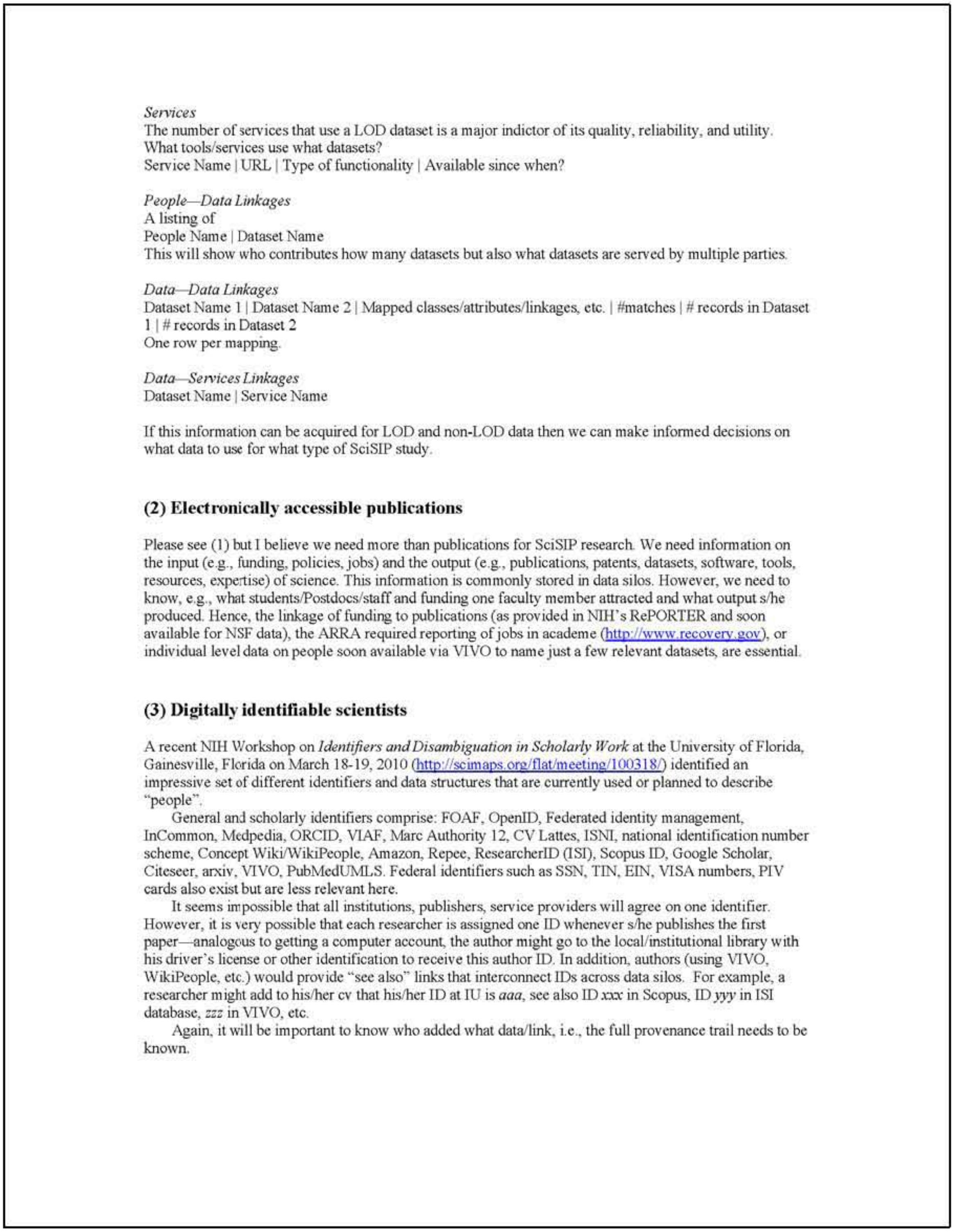
The Scholarly Database (SDB) (http://sdb.slis.indiana.edu) at Indiana University supports federated search over 23,000,000 MEDLINE papers, USPTO patents, NSF awards, and NIH awards. Matching records can be downloaded as csv file but also in network format, e.g., co-author, co-investigator, co-inventor, patent citation networks, and formats suitable for burst analysis in the NWB Tool (http://nwb.slis.indiana.edu) and Sci2 Tool (http://sci.slis.indiana.edu). In April 2010, SDB has more than 220 registered users from 26 countries on 5 continents. We are in the process of exposing the SDB datasets to the Semantic Web as Linked Open Data.

The VIVO National Researcher Network (http://vivoweb.org) will soon expose high quality people data from systems of record (Human Resource, Sponsored Research, Course databases from academic and government institutions) to Linked Open Data.

Linked Open Data (LOD) (http://linkeddata.org) is relevant to this NSF workshop as it makes many datasets openly available in a structured and interlinked form. However, before LOD can be used for SciSIP studies, it needs to be known who is exposing what data semantically, exactly what data is exposed, and what linkages exist. The provenance trail, i.e., what data came from what source and what changes/unifications/linkages were made, needs to be documented in an audible fashion. Below I provide a listing of the kinds of data I/others need to understand together with sample data formats.

People that serve LOD
Name | Institution | Contact info/email | Geolocation (ZIP if in US, city+country otherwise)

Datasets
Dataset Name | Original Source | URL | # Records | Link to raw data sample | Ontology/structure/data dictionary | topic coverage, e.g., medicine, CS | Type, e.g., patents, funding, genes | Available in LOD since when?
Are there also derivative datasets in LOD? For example datasets that add additional (calculated) values or unify names, geolocations, etc?



Services
The number of services that use a LOD dataset is a major indictor of its quality, reliability, and utility.
What tools/services use what datasets?
Service Name | URL | Type of functionality | Available since when?

People—Data Linkages
A listing of
People Name | Dataset Name
This will show who contributes how many datasets but also what datasets are served by multiple parties.

Data—Data Linkages
Dataset Name 1 | Dataset Name 2 | Mapped classes/attributes/linkages, etc. | #matches | # records in Dataset 1 | # records in Dataset 2
One row per mapping.

Data—Services Linkages
Dataset Name | Service Name

If this information can be acquired for LOD and non-LOD data then we can make informed decisions on what data to use for what type of SciSIP study.

(2) Electronically accessible publications

Please see (1) but I believe we need more than publications for SciSIP research. We need information on the input (e.g., funding, policies, jobs) and the output (e.g., publications, patents, datasets, software, tools, resources, expertise) of science. This information is commonly stored in data silos. However, we need to know, e.g., what students/Postdocs/staff and funding one faculty member attracted and what output s/he produced. Hence, the linkage of funding to publications (as provided in NIH's RePORTER and soon available for NSF data), the ARRA required reporting of jobs in academe (http://www.recovery.gov), or individual level data on people soon available via VIVO to name just a few relevant datasets, are essential.

(3) Digitally identifiable scientists

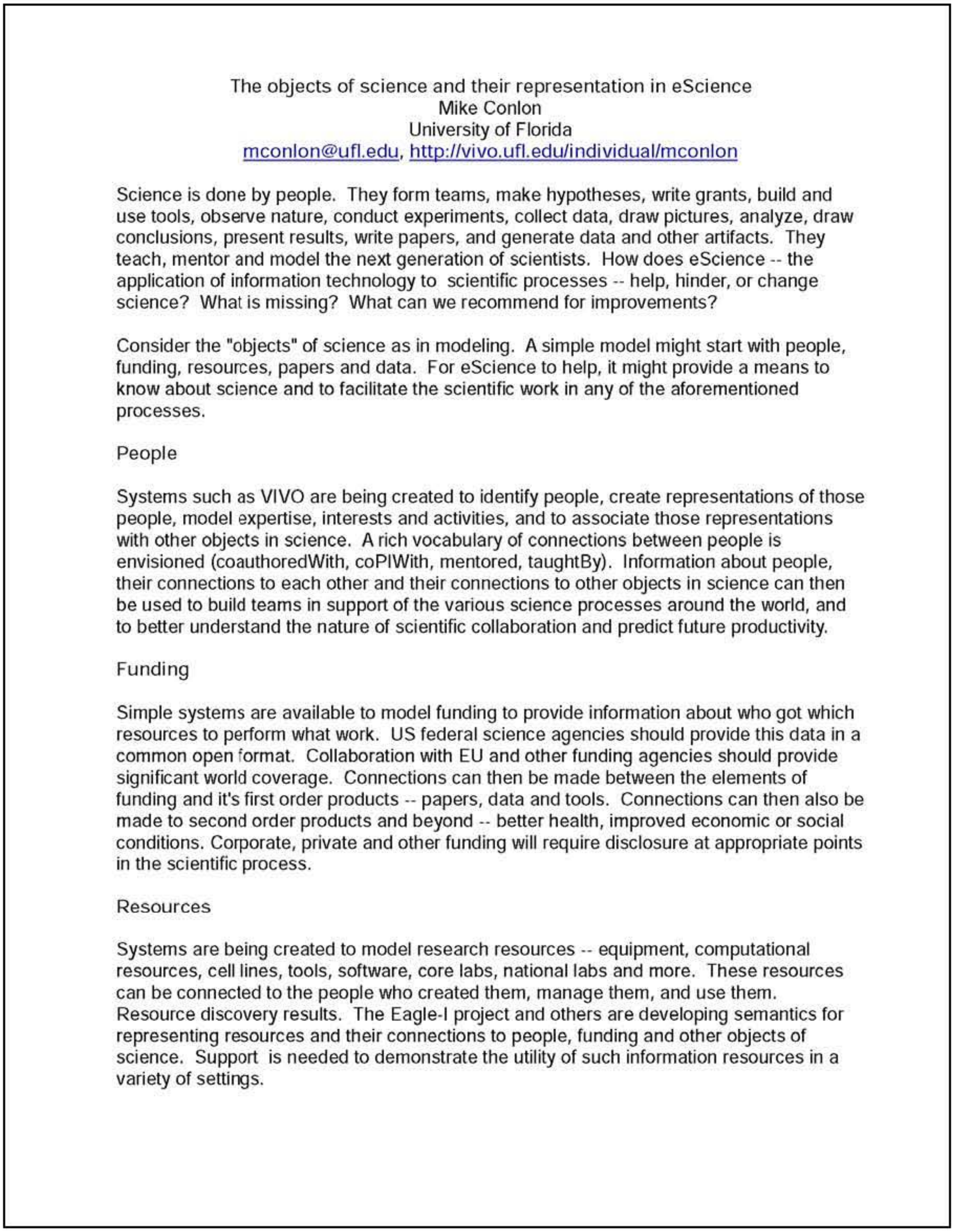
A recent NIH Workshop on *Identifiers and Disambiguation in Scholarly Work* at the University of Florida, Gainesville, Florida on March 18-19, 2010 (http://scimaps.org/flat/meeting/100318/) identified an impressive set of different identifiers and data structures that are currently used or planned to describe "people".

General and scholarly identifiers comprise: FOAF, OpenID, Federated identity management, InCommon, Medpedia, ORCID, VIAF, Marc Authority 12, CV Lattes, ISNI, national identification number scheme, Concept Wiki/WikiPeople, Amazon, Repee, ResearcherID (ISI), Scopus ID, Google Scholar, Citeseer, arxiv, VIVO, PubMedUMLS. Federal identifiers such as SSN, TIN, EIN, VISA numbers, PIV cards also exist but are less relevant here.

It seems impossible that all institutions, publishers, service providers will agree on one identifier. However, it is very possible that each researcher is assigned one ID whenever s/he publishes the first paper—analogous to getting a computer account, the author might go to the local/institutional library with his driver's license or other identification to receive this author ID. In addition, authors (using VIVO, WikiPeople, etc.) would provide "see also" links that interconnect IDs across data silos. For example, a researcher might add to his/her cv that his/her ID at IU is *aaa*, see also ID *xxx* in Scopus, ID *yyy* in ISI database, *zzz* in VIVO, etc.

Again, it will be important to know who added what data/link, i.e., the full provenance trail needs to be known.

Conlon, Mike, "The Objects of Science and Their Representation in eScience"



The objects of science and their representation in eScience

Mike Conlon
University of Florida
mconlon@ufl.edu, http://vivo.ufl.edu/individual/mconlon

Science is done by people. They form teams, make hypotheses, write grants, build and use tools, observe nature, conduct experiments, collect data, draw pictures, analyze, draw conclusions, present results, write papers, and generate data and other artifacts. They teach, mentor and model the next generation of scientists. How does eScience -- the application of information technology to scientific processes -- help, hinder, or change science? What is missing? What can we recommend for improvements?

Consider the "objects" of science as in modeling. A simple model might start with people, funding, resources, papers and data. For eScience to help, it might provide a means to know about science and to facilitate the scientific work in any of the aforementioned processes.

People

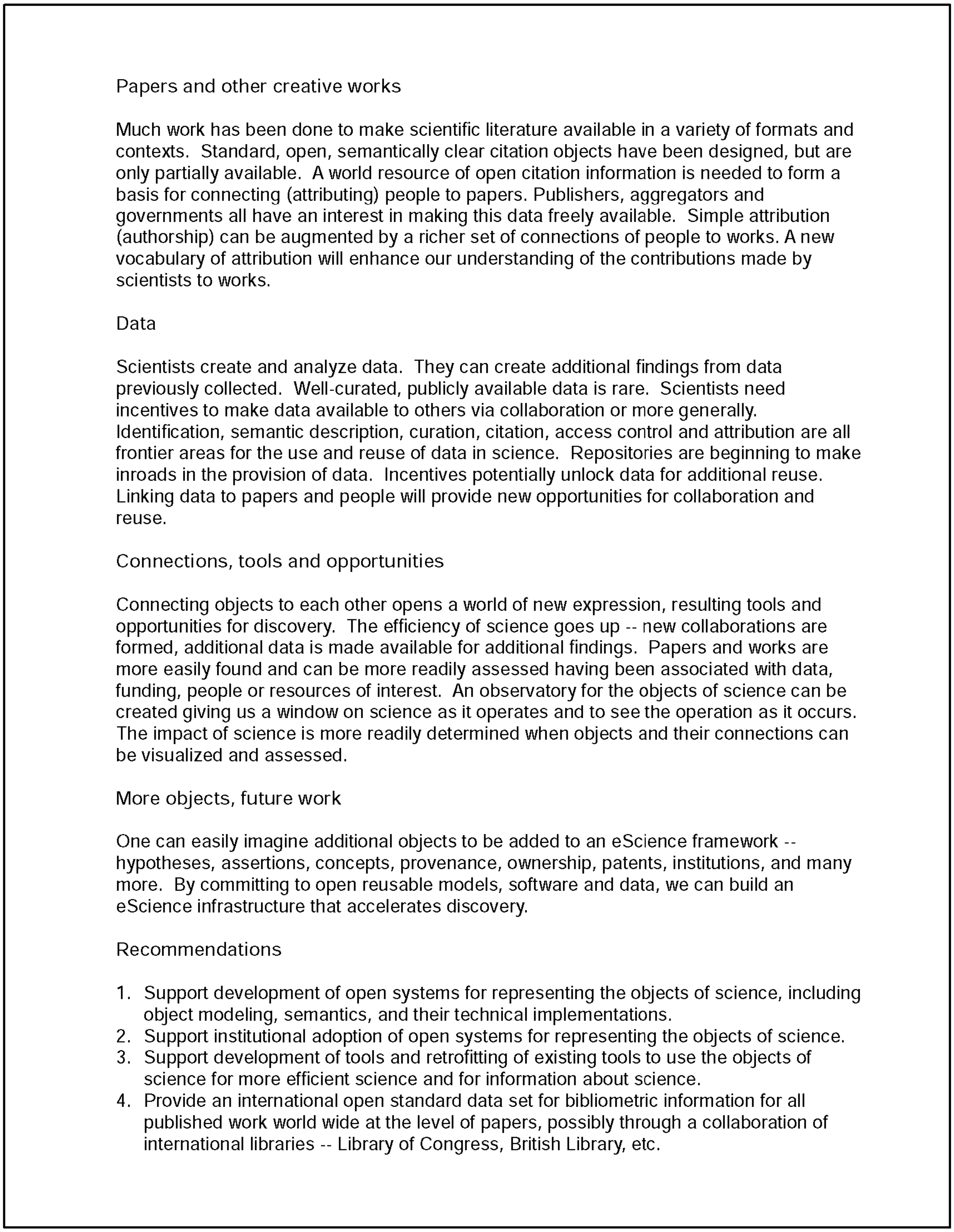
Systems such as VIVO are being created to identify people, create representations of those people, model expertise, interests and activities, and to associate those representations with other objects in science. A rich vocabulary of connections between people is envisioned (coauthoredWith, coPIWith, mentored, taughtBy). Information about people, their connections to each other and their connections to other objects in science can then be used to build teams in support of the various science processes around the world, and to better understand the nature of scientific collaboration and predict future productivity.

Funding

Simple systems are available to model funding to provide information about who got which resources to perform what work. US federal science agencies should provide this data in a common open format. Collaboration with EU and other funding agencies should provide significant world coverage. Connections can then be made between the elements of funding and it's first order products -- papers, data and tools. Connections can then also be made to second order products and beyond -- better health, improved economic or social conditions. Corporate, private and other funding will require disclosure at appropriate points in the scientific process.

Resources

Systems are being created to model research resources -- equipment, computational resources, cell lines, tools, software, core labs, national labs and more. These resources can be connected to the people who created them, manage them, and use them. Resource discovery results. The Eagle-I project and others are developing semantics for representing resources and their connections to people, funding and other objects of science. Support is needed to demonstrate the utility of such information resources in a variety of settings.



Papers and other creative works

Much work has been done to make scientific literature available in a variety of formats and contexts. Standard, open, semantically clear citation objects have been designed, but are only partially available. A world resource of open citation information is needed to form a basis for connecting (attributing) people to papers. Publishers, aggregators and governments all have an interest in making this data freely available. Simple attribution (authorship) can be augmented by a richer set of connections of people to works. A new vocabulary of attribution will enhance our understanding of the contributions made by scientists to works.

Data

Scientists create and analyze data. They can create additional findings from data previously collected. Well-curated, publicly available data is rare. Scientists need incentives to make data available to others via collaboration or more generally. Identification, semantic description, curation, citation, access control and attribution are all frontier areas for the use and reuse of data in science. Repositories are beginning to make inroads in the provision of data. Incentives potentially unlock data for additional reuse. Linking data to papers and people will provide new opportunities for collaboration and reuse.

Connections, tools and opportunities

Connecting objects to each other opens a world of new expression, resulting tools and opportunities for discovery. The efficiency of science goes up -- new collaborations are formed, additional data is made available for additional findings. Papers and works are more easily found and can be more readily assessed having been associated with data, funding, people or resources of interest. An observatory for the objects of science can be created giving us a window on science as it operates and to see the operation as it occurs. The impact of science is more readily determined when objects and their connections can be visualized and assessed.

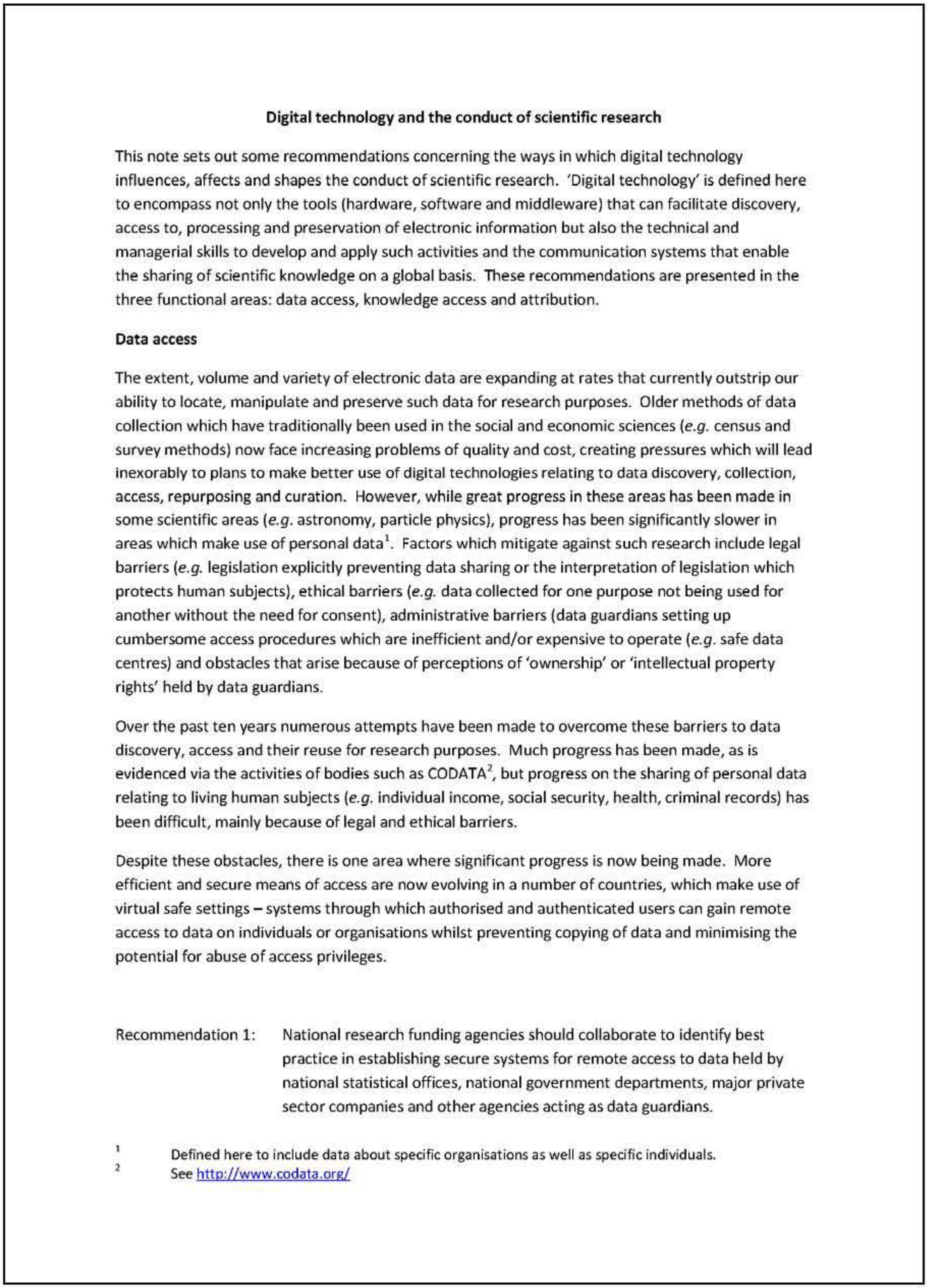
More objects, future work

One can easily imagine additional objects to be added to an eScience framework -- hypotheses, assertions, concepts, provenance, ownership, patents, institutions, and many more. By committing to open reusable models, software and data, we can build an eScience infrastructure that accelerates discovery.

Recommendations

1. Support development of open systems for representing the objects of science, including object modeling, semantics, and their technical implementations.
2. Support institutional adoption of open systems for representing the objects of science.
3. Support development of tools and retrofitting of existing tools to use the objects of science for more efficient science and for information about science.
4. Provide an international open standard data set for bibliometric information for all published work world wide at the level of papers, possibly through a collaboration of international libraries -- Library of Congress, British Library, etc.

Elias, Peter, “Digital Technology and the Conduct of Scientific Research”



Digital technology and the conduct of scientific research

This note sets out some recommendations concerning the ways in which digital technology influences, affects and shapes the conduct of scientific research. ‘Digital technology’ is defined here to encompass not only the tools (hardware, software and middleware) that can facilitate discovery, access to, processing and preservation of electronic information but also the technical and managerial skills to develop and apply such activities and the communication systems that enable the sharing of scientific knowledge on a global basis. These recommendations are presented in the three functional areas: data access, knowledge access and attribution.

Data access

The extent, volume and variety of electronic data are expanding at rates that currently outstrip our ability to locate, manipulate and preserve such data for research purposes. Older methods of data collection which have traditionally been used in the social and economic sciences (*e.g.* census and survey methods) now face increasing problems of quality and cost, creating pressures which will lead inexorably to plans to make better use of digital technologies relating to data discovery, collection, access, repurposing and curation. However, while great progress in these areas has been made in some scientific areas (*e.g.* astronomy, particle physics), progress has been significantly slower in areas which make use of personal data[1]. Factors which mitigate against such research include legal barriers (*e.g.* legislation explicitly preventing data sharing or the interpretation of legislation which protects human subjects), ethical barriers (*e.g.* data collected for one purpose not being used for another without the need for consent), administrative barriers (data guardians setting up cumbersome access procedures which are inefficient and/or expensive to operate (*e.g.* safe data centres) and obstacles that arise because of perceptions of ‘ownership’ or ‘intellectual property rights’ held by data guardians.

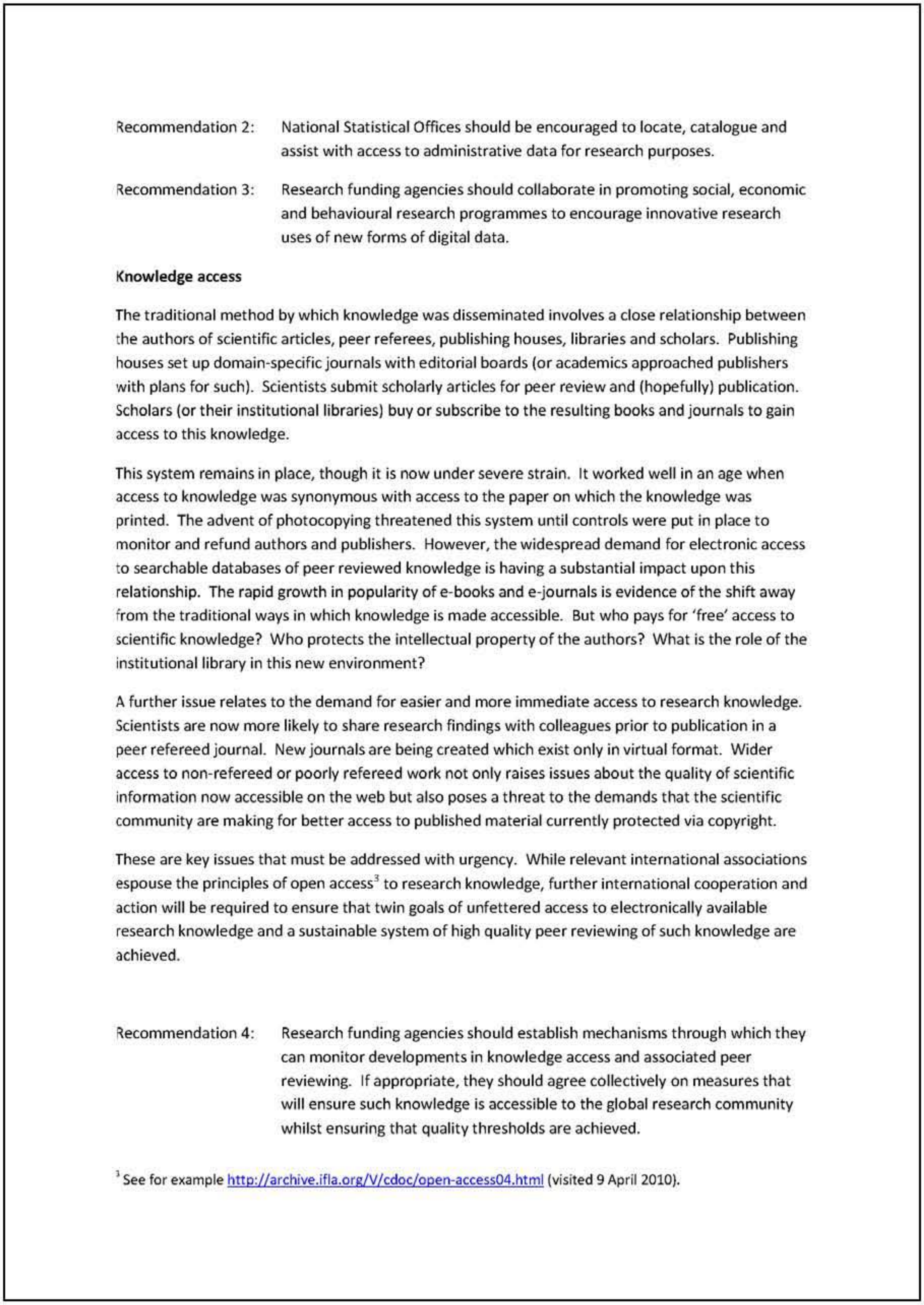
Over the past ten years numerous attempts have been made to overcome these barriers to data discovery, access and their reuse for research purposes. Much progress has been made, as is evidenced via the activities of bodies such as CODATA[2], but progress on the sharing of personal data relating to living human subjects (*e.g.* individual income, social security, health, criminal records) has been difficult, mainly because of legal and ethical barriers.

Despite these obstacles, there is one area where significant progress is now being made. More efficient and secure means of access are now evolving in a number of countries, which make use of virtual safe settings – systems through which authorised and authenticated users can gain remote access to data on individuals or organisations whilst preventing copying of data and minimising the potential for abuse of access privileges.

Recommendation 1: National research funding agencies should collaborate to identify best practice in establishing secure systems for remote access to data held by national statistical offices, national government departments, major private sector companies and other agencies acting as data guardians.

[1] Defined here to include data about specific organisations as well as specific individuals.

[2] See http://www.codata.org/



Recommendation 2: National Statistical Offices should be encouraged to locate, catalogue and assist with access to administrative data for research purposes.

Recommendation 3: Research funding agencies should collaborate in promoting social, economic and behavioural research programmes to encourage innovative research uses of new forms of digital data.

Knowledge access

The traditional method by which knowledge was disseminated involves a close relationship between the authors of scientific articles, peer referees, publishing houses, libraries and scholars. Publishing houses set up domain-specific journals with editorial boards (or academics approached publishers with plans for such). Scientists submit scholarly articles for peer review and (hopefully) publication. Scholars (or their institutional libraries) buy or subscribe to the resulting books and journals to gain access to this knowledge.

This system remains in place, though it is now under severe strain. It worked well in an age when access to knowledge was synonymous with access to the paper on which the knowledge was printed. The advent of photocopying threatened this system until controls were put in place to monitor and refund authors and publishers. However, the widespread demand for electronic access to searchable databases of peer reviewed knowledge is having a substantial impact upon this relationship. The rapid growth in popularity of e-books and e-journals is evidence of the shift away from the traditional ways in which knowledge is made accessible. But who pays for 'free' access to scientific knowledge? Who protects the intellectual property of the authors? What is the role of the institutional library in this new environment?

A further issue relates to the demand for easier and more immediate access to research knowledge. Scientists are now more likely to share research findings with colleagues prior to publication in a peer refereed journal. New journals are being created which exist only in virtual format. Wider access to non-refereed or poorly refereed work not only raises issues about the quality of scientific information now accessible on the web but also poses a threat to the demands that the scientific community are making for better access to published material currently protected via copyright.

These are key issues that must be addressed with urgency. While relevant international associations espouse the principles of open access[3] to research knowledge, further international cooperation and action will be required to ensure that twin goals of unfettered access to electronically available research knowledge and a sustainable system of high quality peer reviewing of such knowledge are achieved.

Recommendation 4: Research funding agencies should establish mechanisms through which they can monitor developments in knowledge access and associated peer reviewing. If appropriate, they should agree collectively on measures that will ensure such knowledge is accessible to the global research community whilst ensuring that quality thresholds are achieved.

[3] See for example http://archive.ifla.org/V/cdoc/open-access04.html (visited 9 April 2010).

Attribution

The third issue which requires global action to resolve relates to the issue of attribution – the ways in which the work of scientists is recognised and accredited. Currently, attribution for research efforts arises primarily through the authorship of scientific papers, books and journals. While scientific disciplines vary in terms of the ways in which authorship is defined and interpreted by others, this system forms the basis of traditional attempts to rate scientists in terms of their productivity, prestige, contribution to knowledge and the impact of their research.

A number of issues are now arising as scientific research moves increasingly from the national to the international sphere. The first relates to the fact that scientists have always been heavily involved in the design, funding, operation and management of large scale research infrastructures. Ranging from multiple array telescopes to genetic databases and major national longitudinal studies, these resources require a major career commitment from research scientists if they are to make their optimal contribution to global scientific enquiry. The rapid growth in the number and scale of this infrastructure means that there are now increasing demands placed on the scientific community to engage in design, funding, and operational functions for major research infrastuctures. While the systems are in place (or are being put in place) to allow international sharing of such facilities for research purposes, the current system for attribution fails to recognise the efforts made by those who help develop, operate, manage and sustain research infrastructure. Without effective attribution for these functions, the future supply of skills, knowledge and expertise required to make them effective could be at risk.

Recommendation 5: Funding agencies should cooperate to undertake a discipline-based review of all current and planned major scientific research infrastructures. This should focus upon their future staffing requirements at the most senior levels, particularly on succession planning. In the light of such evidence, funding agencies may wish to establish internationally agreed procedures through which the scientific contributions made by senior scientists engaged in infrastructure management and direction can be attributed to them.

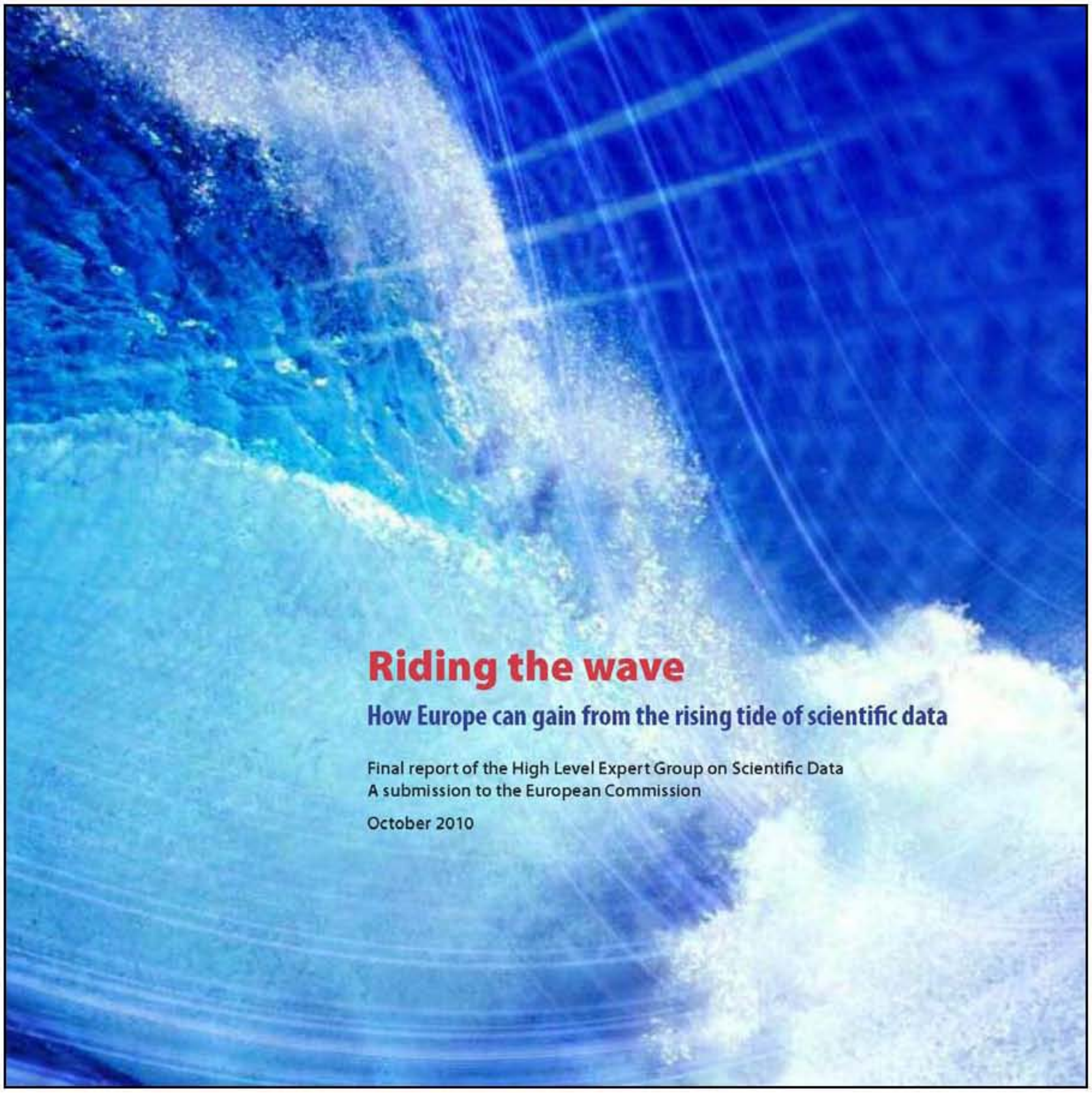
A further issue derives from the demands for open access to research knowledge discussed under the preceding heading. If the quality of electronically available research outputs cannot be effectively controlled and interpreted in an international context, attribution for research outputs is further jeopardised (see recommendation 4).

Peter Elias

Strategic Advisor (Data Resources)
UK Economic and Social Research Council
(Peter.Elias@warwick.ac.uk)

9th April 2010

European Union, "Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data"

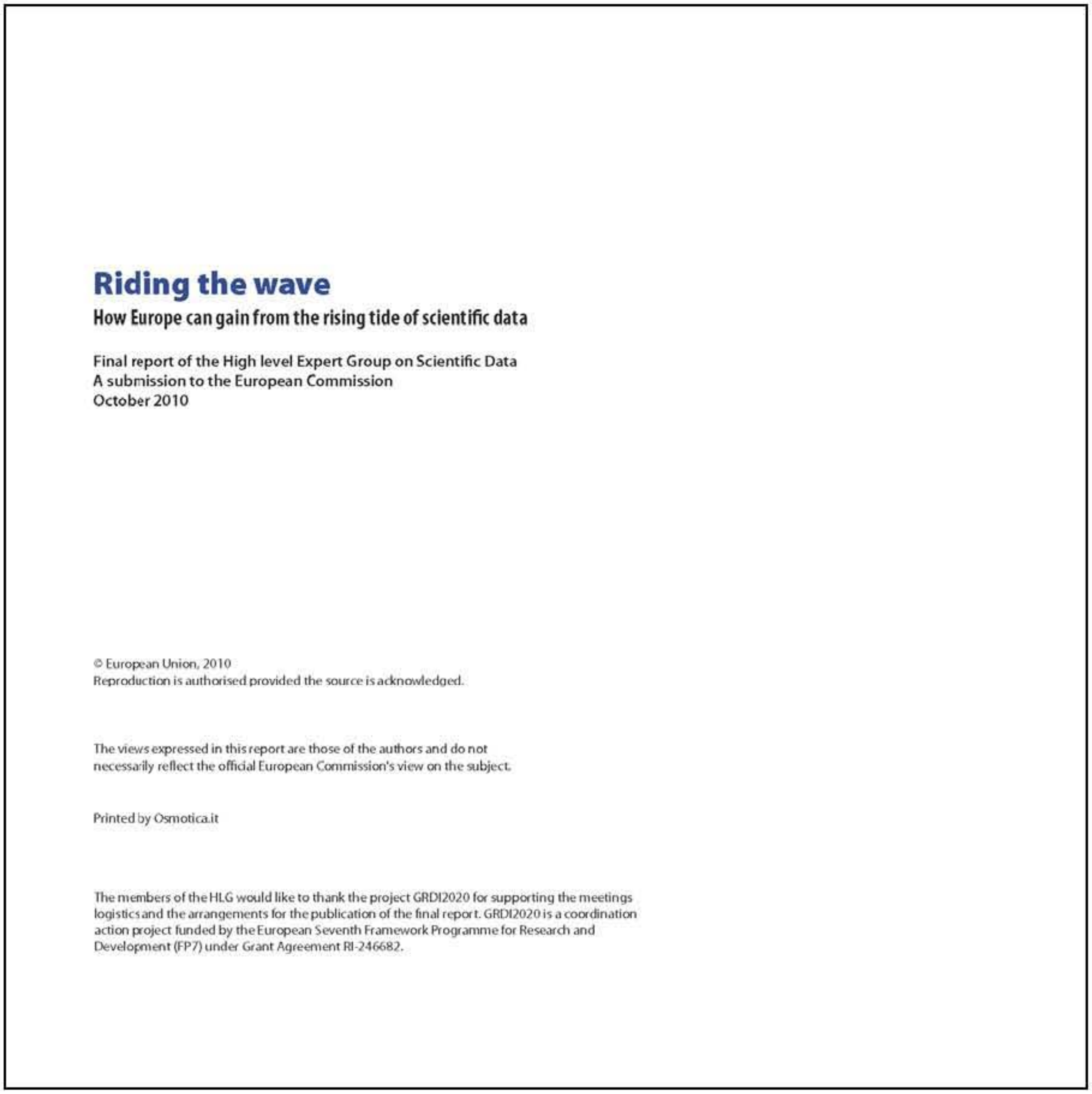


Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission

October 2010



Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High level Expert Group on Scientific Data
A submission to the European Commission
October 2010

The views expressed in this report are those of the authors and do not necessarily reflect the official European Commission's view on the subject.

Printed by Osmotica.it

The members of the HLG would like to thank the project GRDI2020 for supporting the meetings logistics and the arrangements for the publication of the final report. GRDI2020 is a coordination action project funded by the European Seventh Framework Programme for Research and Development (FP7) under Grant Agreement RI-246682.

Digital Agenda
1001100101011101110000100
for Europe
"Ev pean Digital"
Neelie Kroes

Unlocking the full value of scientific data

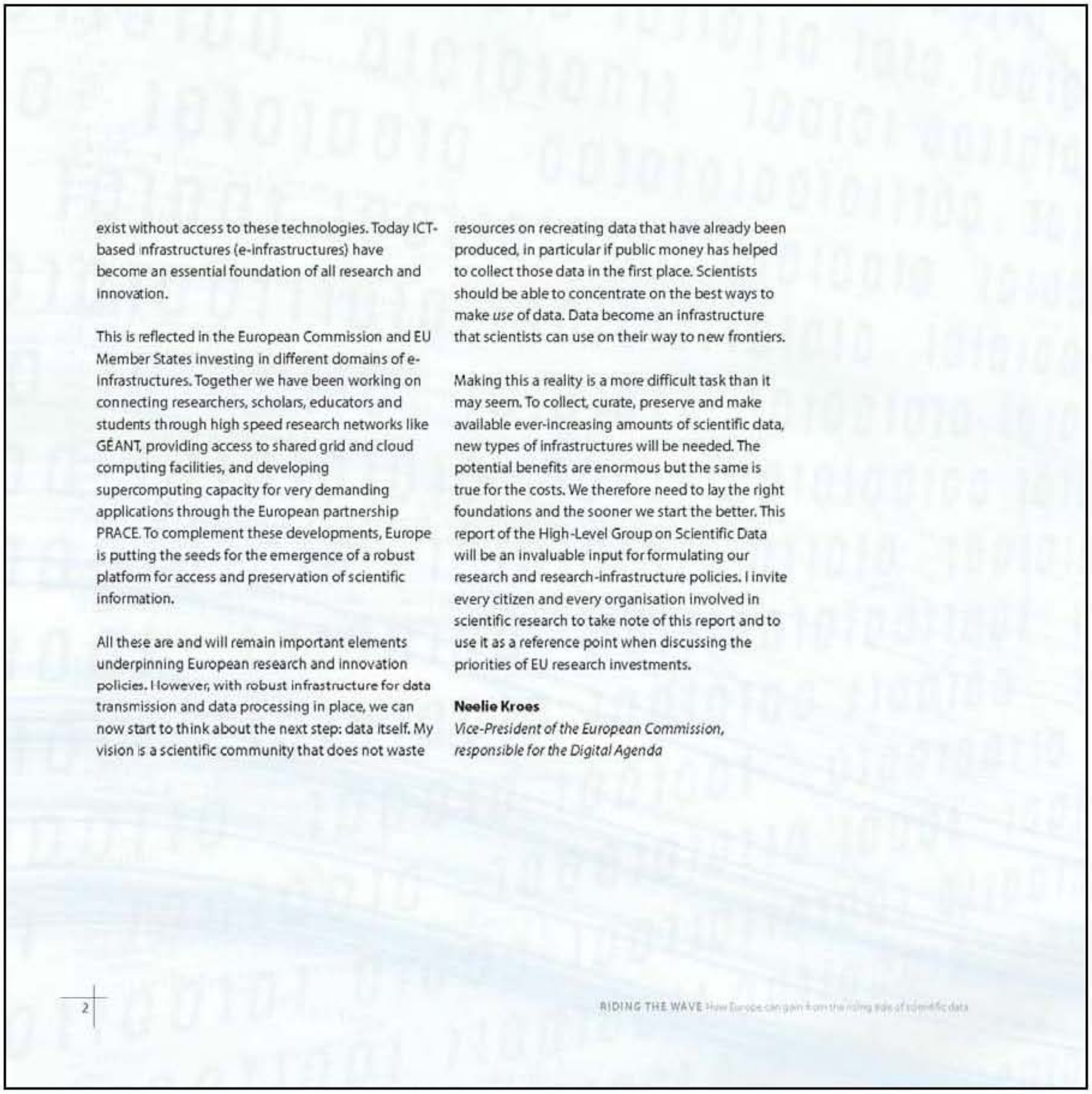
> "Information and Communication Technologies (ICT) are the most recent transformational factors in science."

The Digital Agenda for Europe outlines policies and actions to maximise the benefit of the digital revolution for all. Supporting research and innovation is a key priority of the Agenda, essential if we want to establish a flourishing digital economy by 2020.

Scientific research is supported by its infrastructures: technical tools and instruments and socio-economic systems for organising and sharing knowledge. These have been in constant change for many centuries reflecting advances in technology and change in political systems. Key inventions like the microscope or the telescope resulted in huge scientific progress by allowing the validation or rejection of theories; and the invention of book printing in the 15th century and the organisation of knowledge in research libraries allowed unprecedented access to knowledge.

Information and Communication Technologies (ICT) are the most recent transformational factors in science. They enable close and almost instantaneous collaboration between scientists all over the world and they provide access to unprecedented volumes of scientific information that can in turn be processed on powerful computational platforms. Many younger scientific disciplines would not even

RIDING THE WAVE How Europe can gain from the rising tide of scientific data 1



exist without access to these technologies. Today ICT-based infrastructures (e-infrastructures) have become an essential foundation of all research and innovation.

This is reflected in the European Commission and EU Member States investing in different domains of e-infrastructures. Together we have been working on connecting researchers, scholars, educators and students through high speed research networks like GÉANT, providing access to shared grid and cloud computing facilities, and developing supercomputing capacity for very demanding applications through the European partnership PRACE. To complement these developments, Europe is putting the seeds for the emergence of a robust platform for access and preservation of scientific information.

All these are and will remain important elements underpinning European research and innovation policies. However, with robust infrastructure for data transmission and data processing in place, we can now start to think about the next step: data itself. My vision is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make *use* of data. Data become an infrastructure that scientists can use on their way to new frontiers.

Making this a reality is a more difficult task than it may seem. To collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed. The potential benefits are enormous but the same is true for the costs. We therefore need to lay the right foundations and the sooner we start the better. This report of the High-Level Group on Scientific Data will be an invaluable input for formulating our research and research-infrastructure policies. I invite every citizen and every organisation involved in scientific research to take note of this report and to use it as a reference point when discussing the priorities of EU research investments.

Neelie Kroes
Vice-President of the European Commission,
responsible for the Digital Agenda

2 RIDING THE WAVE How Europe can gain from the rising tide of scientific data

FROM THE CHAIR

On the challenges ahead

I present the report of the High Level Group on the future of scientific data. The importance of facing up to the challenges before us is crucial if European research is to remain at the leading edge globally.

The resulting actions that we propose will affect all areas of research, not just big science. This range has been reflected in the group as we have considered the impact on, for example, the humanities, publishing, and bio-diversity in addition to large international science facilities. Indeed, getting it right will affect the way research is done in the future and will be instrumental in ensuring that the challenges before us are solved in a holistic way rather than allowing individual disciplines to dig entrenched positions. Just how students will be trained in the future, or how the profession of "data scientist" will be developed, are among the questions the resolution of which is still evolving and will present intellectual challenges for both privately and publicly supported research. Critical to everything is how trust can not only be fostered but ensured so that the "Fifth Freedom of Knowledge" is pursued with vigour for the good of all society.

In addition to the High Level Group coming from a diversity of backgrounds, the liveliness of the discussions and the working atmosphere have been a delight and I thank the members for their excellent contributions. Also my thanks to the Commission staff who have entered into the debate with an exemplary degree of open-mindedness. Finally I would like to acknowledge the assistance of the various people who came to the group to share their thoughts and experience with us from around the world, to rapporteur David Giaretta who brought the discussions together into a coherent structure and action plan, and to Richard Hudson who miraculously took our stream of consciousness ideas and turned them into a coherent report.

John Wood
Chair

EXECUTIVE SUMMARY

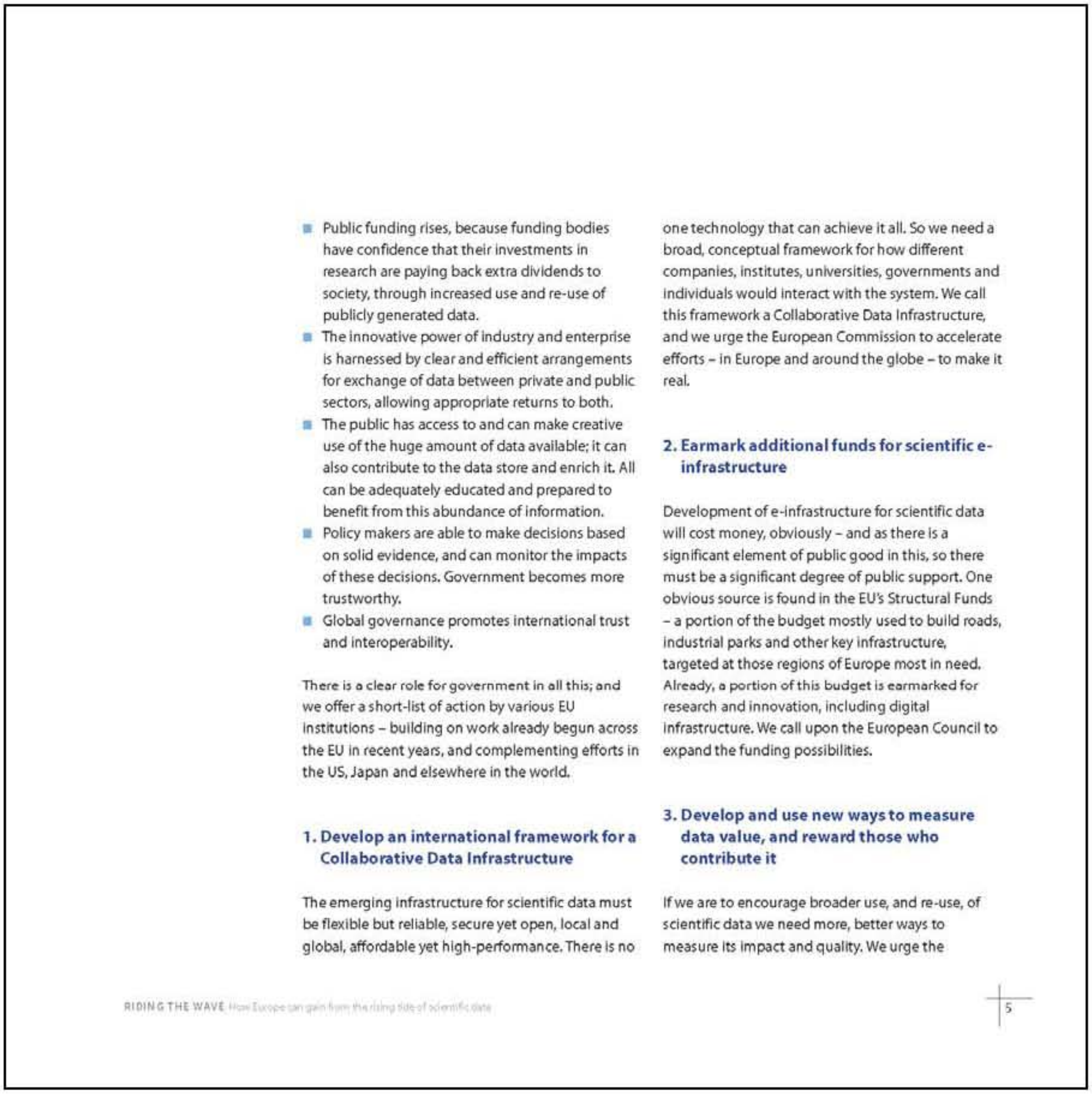
A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge. This report, prepared for the European Commission's Directorate-General for Information Society and Media, identifies the benefits and costs of accelerating the development of a fully functional e-infrastructure for scientific data – a system already emerging piecemeal and spontaneously across the globe, but now in need of a far-seeing, global framework. The outcome will be a vital scientific asset: flexible, reliable, efficient, cross-disciplinary and cross-border.

The benefits are broad. With a proper scientific e-infrastructure, researchers in different domains can collaborate on the same data set, finding new insights. They can share a data set easily across the globe, but also protect its integrity and ownership. They can use, re-use and combine data, increasing productivity. They can more easily solve today's Grand Challenges, such as climate change and energy supply. Indeed, they can engage in whole new forms of scientific inquiry, made possible by the unimaginable power of the e-infrastructure to find correlations, draw inferences and trade ideas and information at a scale we are only beginning to see. For society as a whole, this is beneficial. It empowers amateurs to contribute more easily to the scientific process, politicians to govern more effectively with solid evidence, and the European and global economy to expand.

But there are many challenges. How can we organise such a fiendishly complicated global effort, without hindering its flexibility and openness? How do we incentivise researchers, companies, and individuals to contribute their own data to the e-infrastructure – while still trusting that they can protect their privacy or ownership? How can we manage to preserve all this data, despite changing technologies and needs? How to convey the context and provenance of the data? How to pay for it all?

Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance. Our vision is that, by 2030:

- All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.
- Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.
- Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.



- Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.
- The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.
- The public has access to and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it. All can be adequately educated and prepared to benefit from this abundance of information.
- Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.
- Global governance promotes international trust and interoperability.

There is a clear role for government in all this; and we offer a short-list of action by various EU institutions – building on work already begun across the EU in recent years, and complementing efforts in the US, Japan and elsewhere in the world.

1. Develop an international framework for a Collaborative Data Infrastructure

The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance. There is no one technology that can achieve it all. So we need a broad, conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system. We call this framework a Collaborative Data Infrastructure, and we urge the European Commission to accelerate efforts – in Europe and around the globe – to make it real.

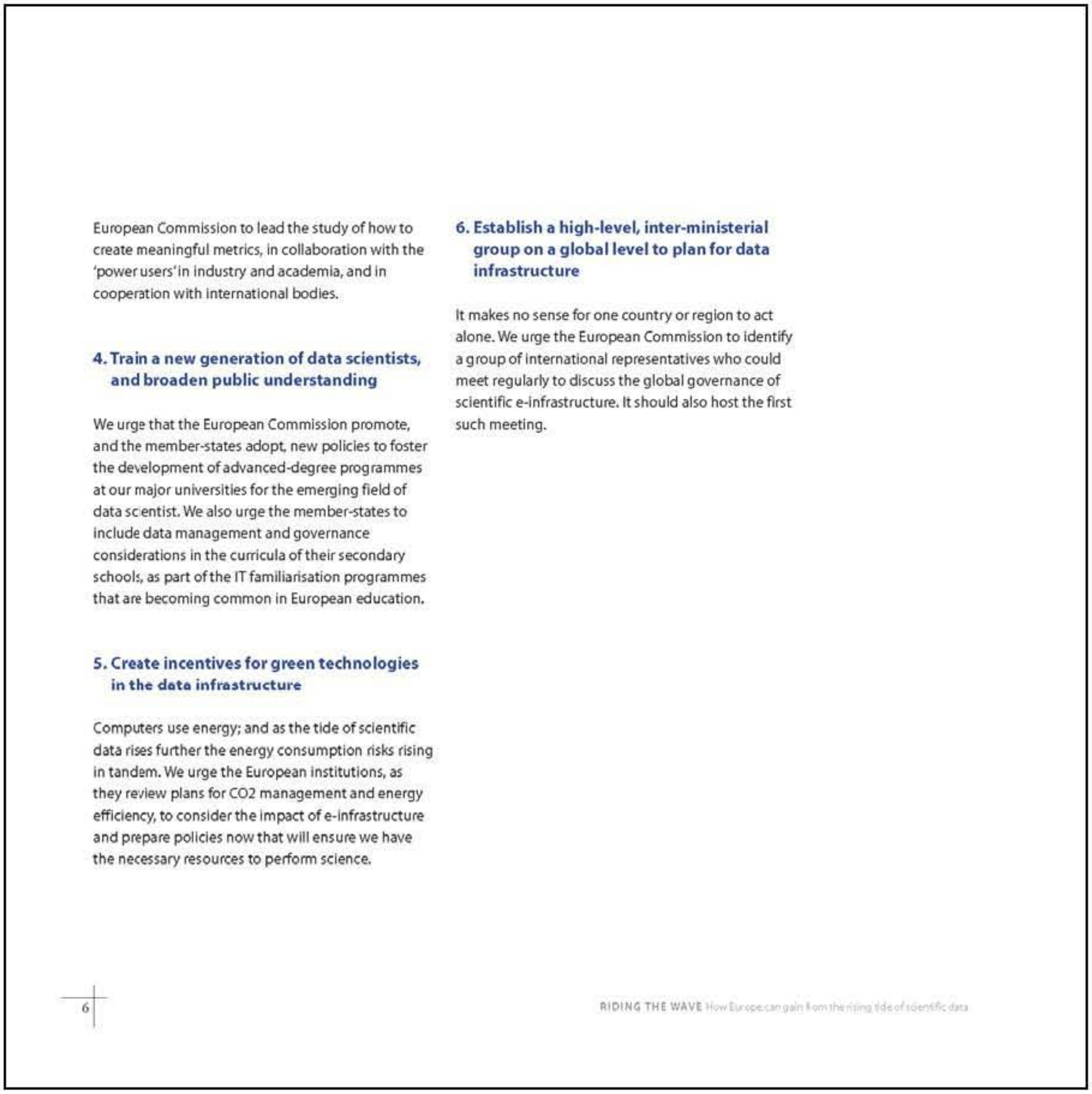
2. Earmark additional funds for scientific e-infrastructure

Development of e-infrastructure for scientific data will cost money, obviously – and as there is a significant element of public good in this, so there must be a significant degree of public support. One obvious source is found in the EU's Structural Funds – a portion of the budget mostly used to build roads, industrial parks and other key infrastructure, targeted at those regions of Europe most in need. Already, a portion of this budget is earmarked for research and innovation, including digital infrastructure. We call upon the European Council to expand the funding possibilities.

3. Develop and use new ways to measure data value, and reward those who contribute it

If we are to encourage broader use, and re-use, of scientific data we need more, better ways to measure its impact and quality. We urge the

RIDING THE WAVE How Europe can gain from the rising tide of scientific data 5



European Commission to lead the study of how to create meaningful metrics, in collaboration with the 'power users' in industry and academia, and in cooperation with international bodies.

4. Train a new generation of data scientists, and broaden public understanding

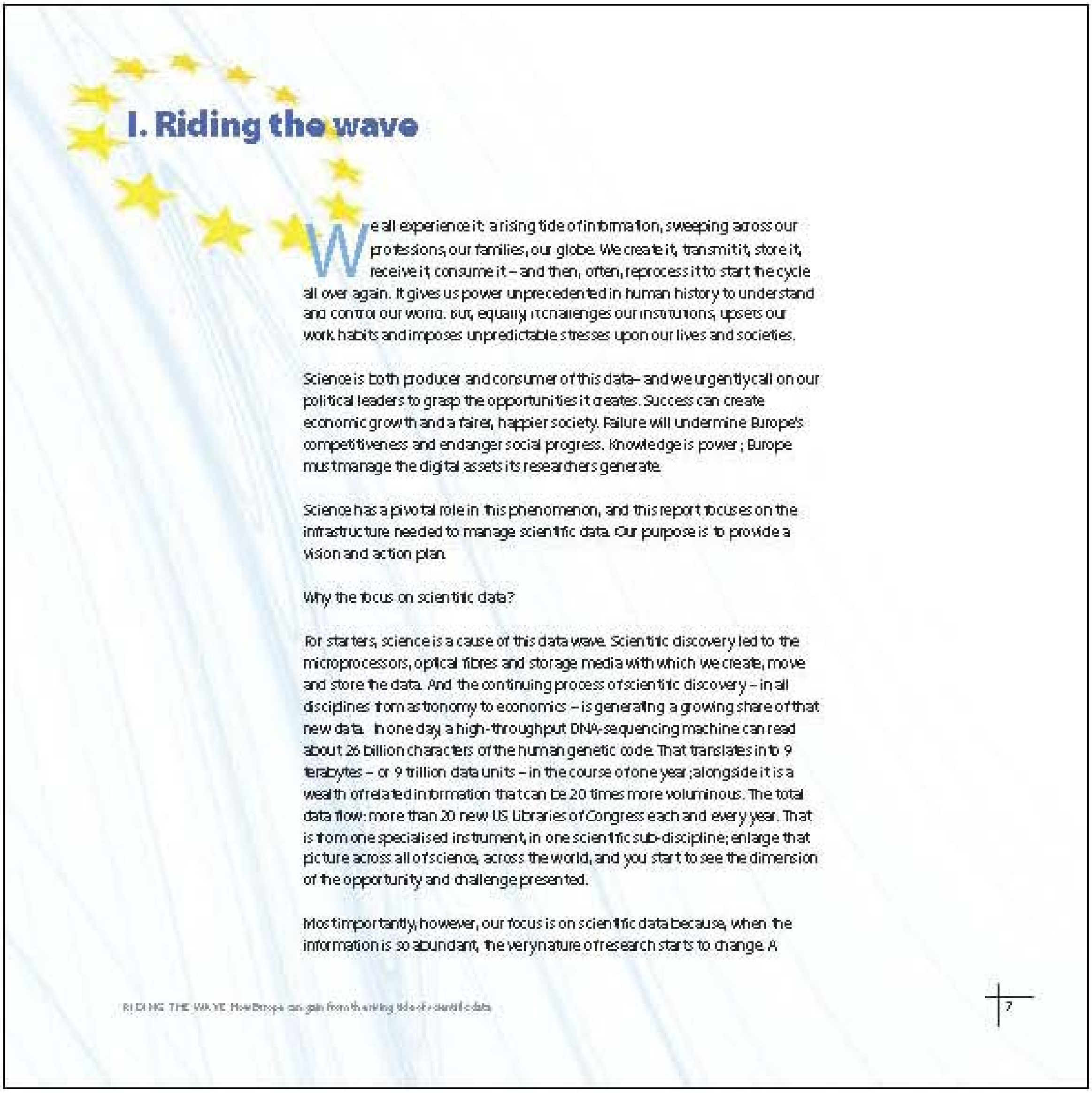
We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for the emerging field of data scientist. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education.

5. Create incentives for green technologies in the data infrastructure

Computers use energy; and as the tide of scientific data rises further the energy consumption risks rising in tandem. We urge the European institutions, as they review plans for CO2 management and energy efficiency, to consider the impact of e-infrastructure and prepare policies now that will ensure we have the necessary resources to perform science.

6. Establish a high-level, inter-ministerial group on a global level to plan for data infrastructure

It makes no sense for one country or region to act alone. We urge the European Commission to identify a group of international representatives who could meet regularly to discuss the global governance of scientific e-infrastructure. It should also host the first such meeting.



I. Riding the wave

We all experience it: a rising tide of information, sweeping across our professions, our families, our globe. We create it, transmit it, store it, receive it, consume it – and then, often, reprocess it to start the cycle all over again. It gives us power unprecedented in human history to understand and control our world. But, equally, it challenges our institutions, upsets our work habits and imposes unpredictable stresses upon our lives and societies.

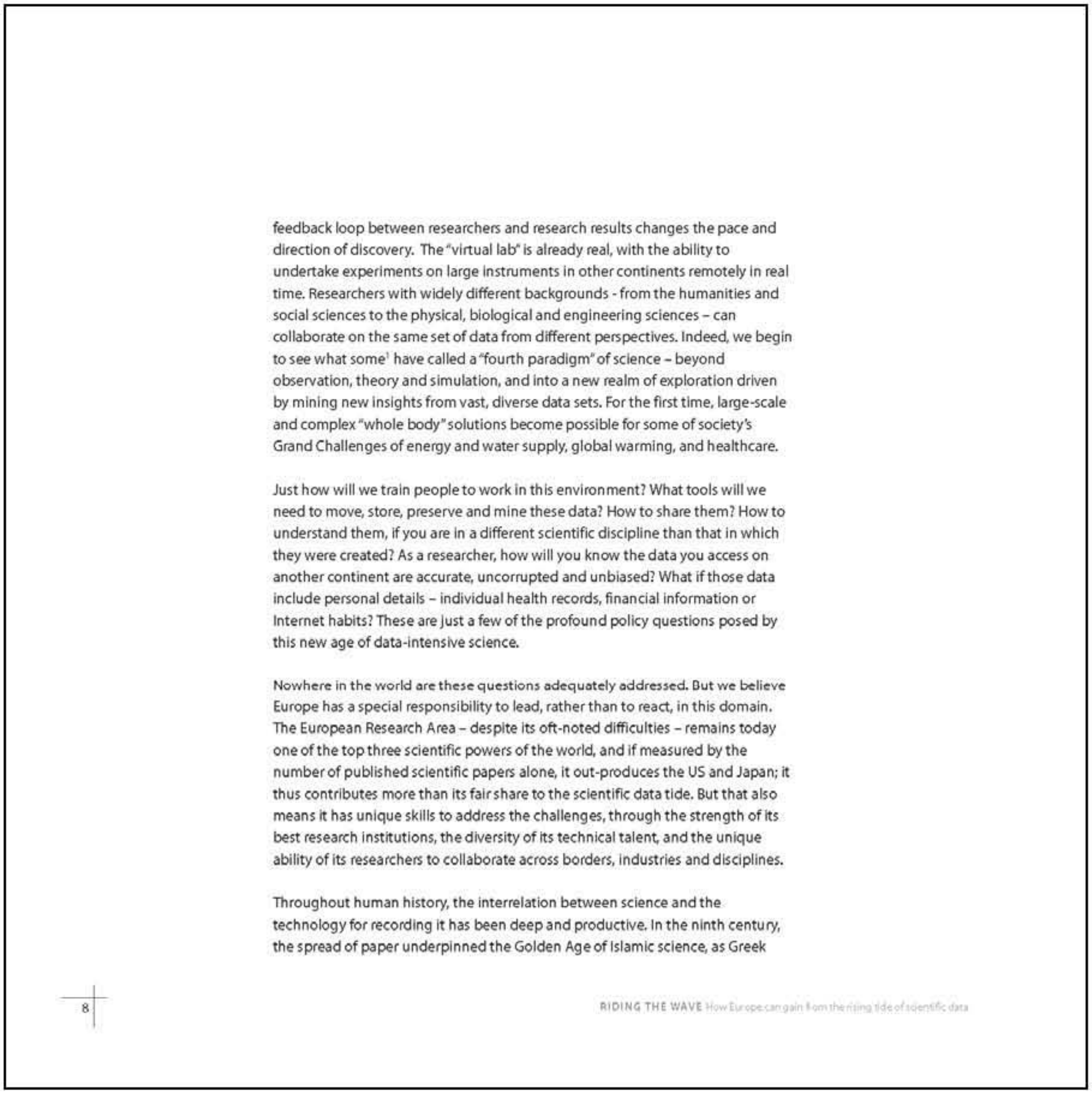
Science is both producer and consumer of this data– and we urgently call on our political leaders to grasp the opportunities it creates. Success can create economic growth and a fairer, happier society. Failure will undermine Europe's competitiveness and endanger social progress. Knowledge is power ; Europe must manage the digital assets its researchers generate.

Science has a pivotal role in this phenomenon, and this report focuses on the infrastructure needed to manage scientific data. Our purpose is to provide a vision and action plan.

Why the focus on scientific data?

For starters, science is a cause of this data wave. Scientific discovery led to the microprocessors, optical fibres and storage media with which we create, move and store the data. And the continuing process of scientific discovery – in all disciplines from astronomy to economics – is generating a growing share of that new data. In one day a high-throughput DNA-sequencing machine can read about 26 billion characters of the human genetic code. That translates into 9 terabytes – or 9 trillion data units – in the course of one year ; alongside it is a wealth of related information that can be 20 times more voluminous. The total data flow: more than 20 new US Libraries of Congress each and every year. That is from one specialised instrument, in one scientific sub-discipline; enlarge that picture across all of science, across the world, and you start to see the dimension of the opportunity and challenge presented.

Most importantly, however, our focus is on scientific data because, when the information is so abundant, the very nature of research starts to change. A

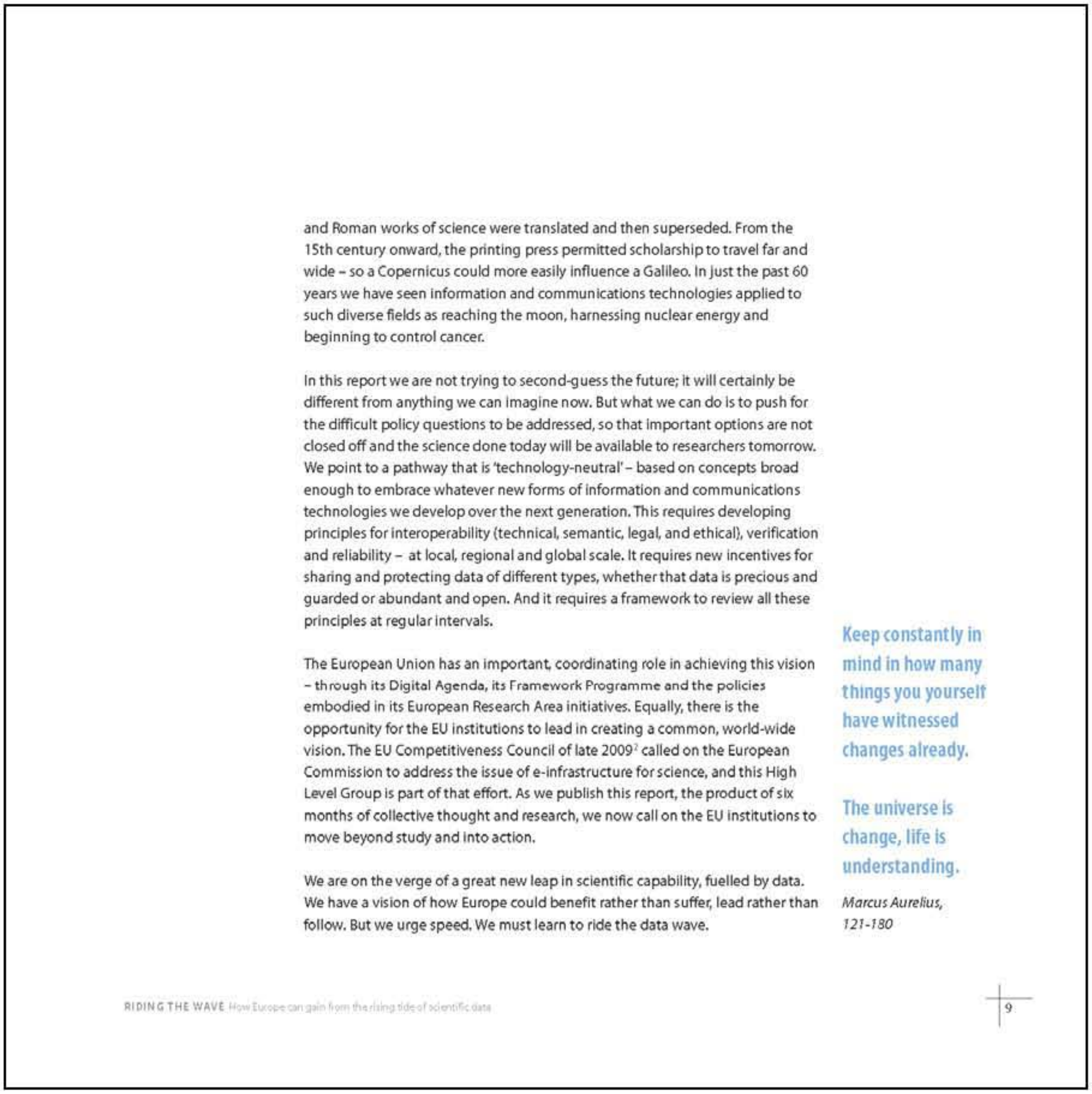


feedback loop between researchers and research results changes the pace and direction of discovery. The "virtual lab" is already real, with the ability to undertake experiments on large instruments in other continents remotely in real time. Researchers with widely different backgrounds - from the humanities and social sciences to the physical, biological and engineering sciences – can collaborate on the same set of data from different perspectives. Indeed, we begin to see what some[1] have called a "fourth paradigm" of science – beyond observation, theory and simulation, and into a new realm of exploration driven by mining new insights from vast, diverse data sets. For the first time, large-scale and complex "whole body" solutions become possible for some of society's Grand Challenges of energy and water supply, global warming, and healthcare.

Just how will we train people to work in this environment? What tools will we need to move, store, preserve and mine these data? How to share them? How to understand them, if you are in a different scientific discipline than that in which they were created? As a researcher, how will you know the data you access on another continent are accurate, uncorrupted and unbiased? What if those data include personal details – individual health records, financial information or Internet habits? These are just a few of the profound policy questions posed by this new age of data-intensive science.

Nowhere in the world are these questions adequately addressed. But we believe Europe has a special responsibility to lead, rather than to react, in this domain. The European Research Area – despite its oft-noted difficulties – remains today one of the top three scientific powers of the world, and if measured by the number of published scientific papers alone, it out-produces the US and Japan; it thus contributes more than its fair share to the scientific data tide. But that also means it has unique skills to address the challenges, through the strength of its best research institutions, the diversity of its technical talent, and the unique ability of its researchers to collaborate across borders, industries and disciplines.

Throughout human history, the interrelation between science and the technology for recording it has been deep and productive. In the ninth century, the spread of paper underpinned the Golden Age of Islamic science, as Greek



and Roman works of science were translated and then superseded. From the 15th century onward, the printing press permitted scholarship to travel far and wide – so a Copernicus could more easily influence a Galileo. In just the past 60 years we have seen information and communications technologies applied to such diverse fields as reaching the moon, harnessing nuclear energy and beginning to control cancer.

In this report we are not trying to second-guess the future; it will certainly be different from anything we can imagine now. But what we can do is to push for the difficult policy questions to be addressed, so that important options are not closed off and the science done today will be available to researchers tomorrow. We point to a pathway that is 'technology-neutral' – based on concepts broad enough to embrace whatever new forms of information and communications technologies we develop over the next generation. This requires developing principles for interoperability (technical, semantic, legal, and ethical), verification and reliability – at local, regional and global scale. It requires new incentives for sharing and protecting data of different types, whether that data is precious and guarded or abundant and open. And it requires a framework to review all these principles at regular intervals.

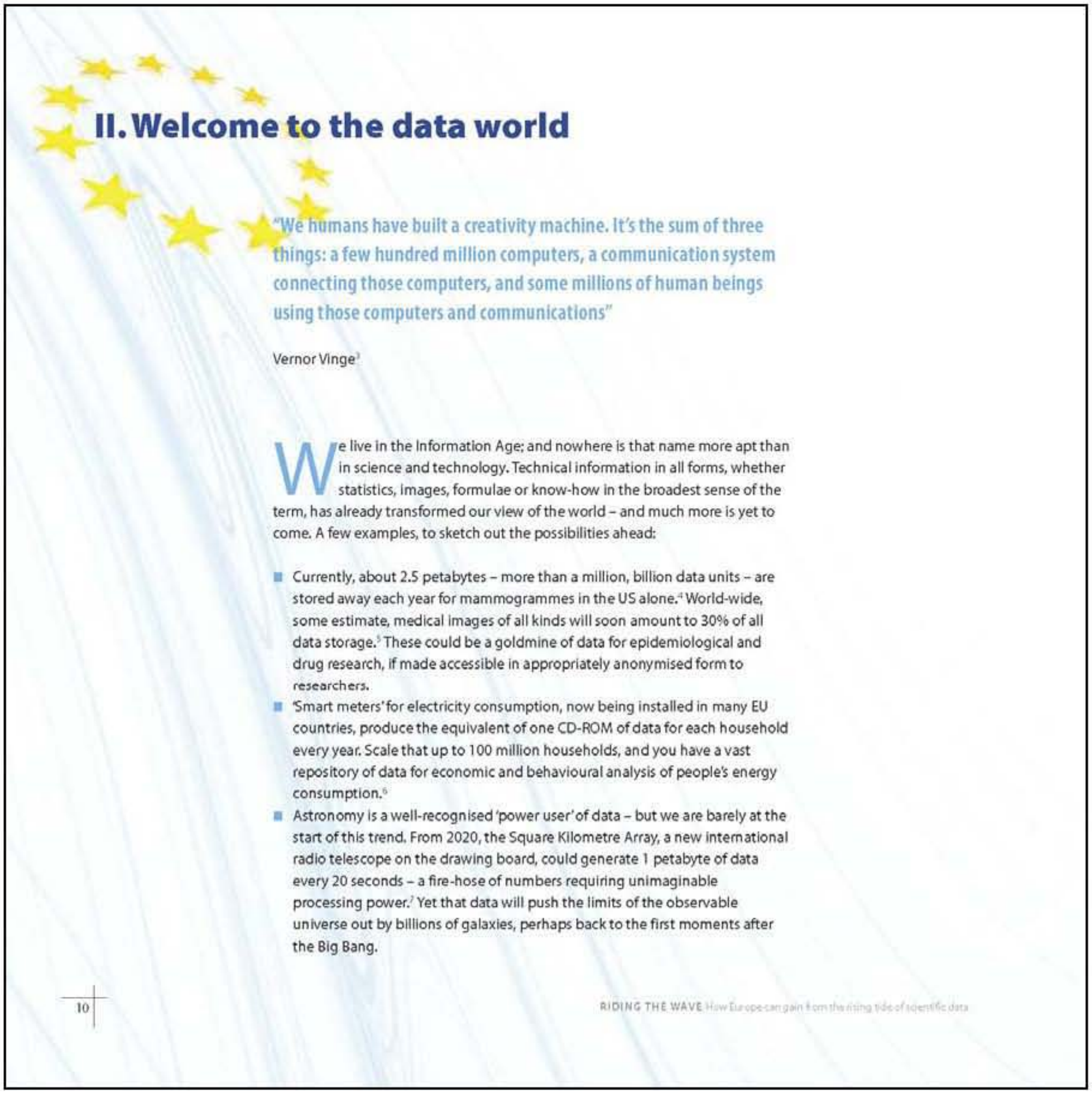
The European Union has an important, coordinating role in achieving this vision – through its Digital Agenda, its Framework Programme and the policies embodied in its European Research Area initiatives. Equally, there is the opportunity for the EU institutions to lead in creating a common, world-wide vision. The EU Competitiveness Council of late 2009[2] called on the European Commission to address the issue of e-infrastructure for science, and this High Level Group is part of that effort. As we publish this report, the product of six months of collective thought and research, we now call on the EU institutions to move beyond study and into action.

We are on the verge of a great new leap in scientific capability, fuelled by data. We have a vision of how Europe could benefit rather than suffer, lead rather than follow. But we urge speed. We must learn to ride the data wave.

Keep constantly in mind in how many things you yourself have witnessed changes already.

The universe is change, life is understanding.

Marcus Aurelius, 121-180



II. Welcome to the data world

"We humans have built a creativity machine. It's the sum of three things: a few hundred million computers, a communication system connecting those computers, and some millions of human beings using those computers and communications"

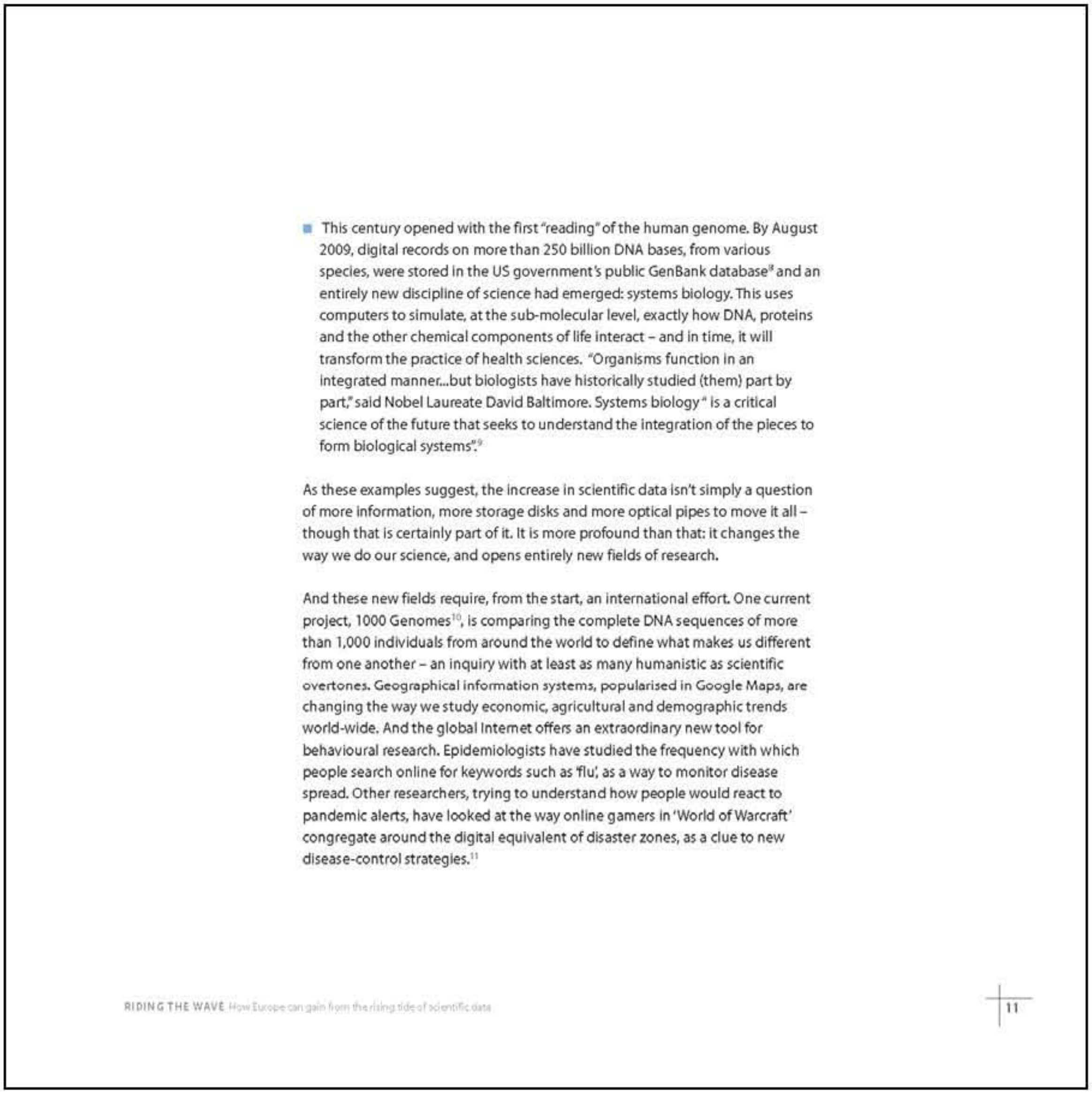
Vernor Vinge[3]

We live in the Information Age; and nowhere is that name more apt than in science and technology. Technical information in all forms, whether statistics, images, formulae or know-how in the broadest sense of the term, has already transformed our view of the world – and much more is yet to come. A few examples, to sketch out the possibilities ahead:

- Currently, about 2.5 petabytes – more than a million, billion data units – are stored away each year for mammogrammes in the US alone.[4] World-wide, some estimate, medical images of all kinds will soon amount to 30% of all data storage.[5] These could be a goldmine of data for epidemiological and drug research, if made accessible in appropriately anonymised form to researchers.
- 'Smart meters' for electricity consumption, now being installed in many EU countries, produce the equivalent of one CD-ROM of data for each household every year. Scale that up to 100 million households, and you have a vast repository of data for economic and behavioural analysis of people's energy consumption.[6]
- Astronomy is a well-recognised 'power user' of data – but we are barely at the start of this trend. From 2020, the Square Kilometre Array, a new international radio telescope on the drawing board, could generate 1 petabyte of data every 20 seconds – a fire-hose of numbers requiring unimaginable processing power.[7] Yet that data will push the limits of the observable universe out by billions of galaxies, perhaps back to the first moments after the Big Bang.

10

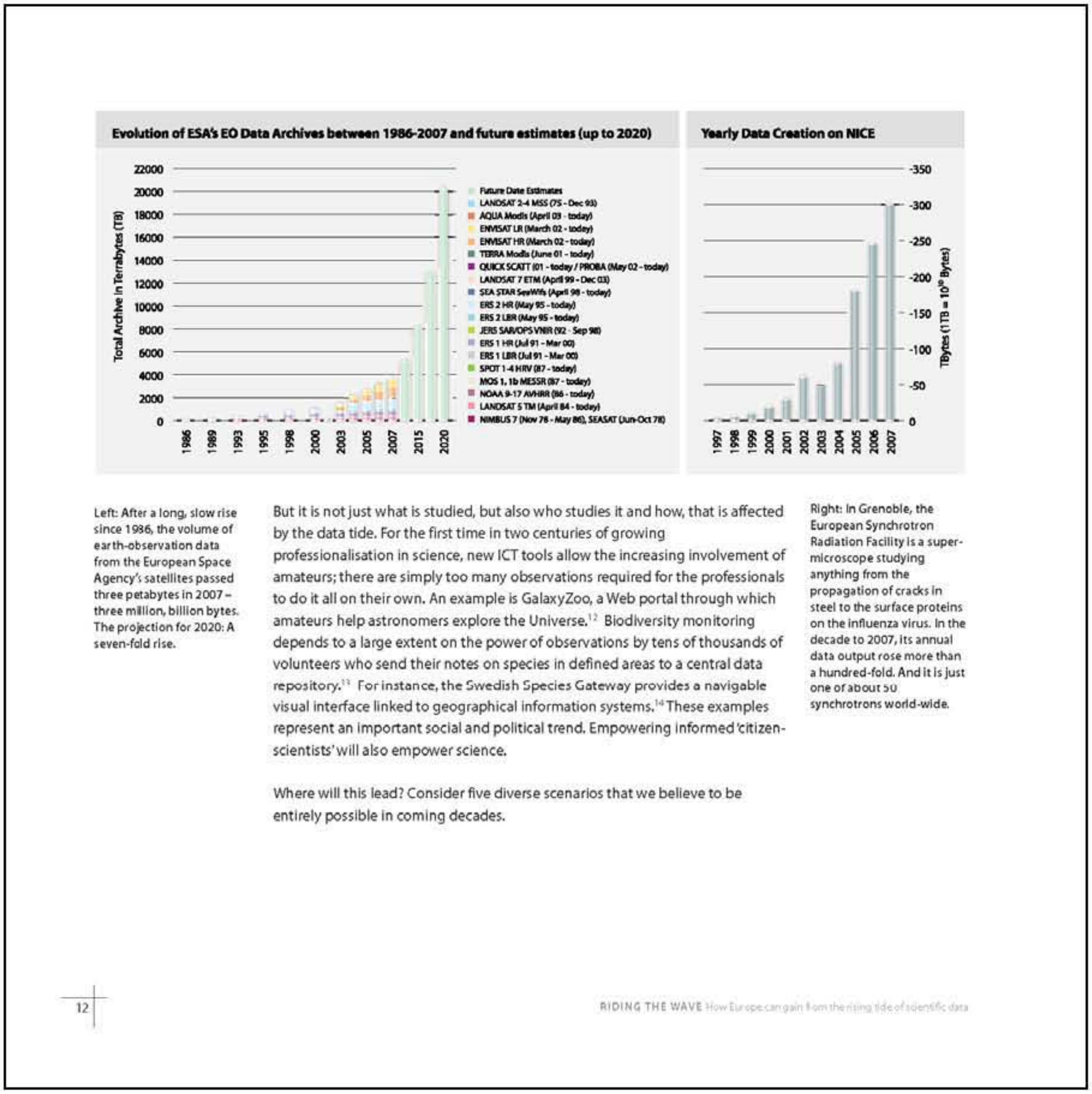
RIDING THE WAVE How Europe can gain from the rising tide of scientific data



- This century opened with the first "reading" of the human genome. By August 2009, digital records on more than 250 billion DNA bases, from various species, were stored in the US government's public GenBank database[8] and an entirely new discipline of science had emerged: systems biology. This uses computers to simulate, at the sub-molecular level, exactly how DNA, proteins and the other chemical components of life interact – and in time, it will transform the practice of health sciences. "Organisms function in an integrated manner...but biologists have historically studied (them) part by part," said Nobel Laureate David Baltimore. Systems biology " is a critical science of the future that seeks to understand the integration of the pieces to form biological systems".[9]

As these examples suggest, the increase in scientific data isn't simply a question of more information, more storage disks and more optical pipes to move it all – though that is certainly part of it. It is more profound than that: it changes the way we do our science, and opens entirely new fields of research.

And these new fields require, from the start, an international effort. One current project, 1000 Genomes[10], is comparing the complete DNA sequences of more than 1,000 individuals from around the world to define what makes us different from one another – an inquiry with at least as many humanistic as scientific overtones. Geographical information systems, popularised in Google Maps, are changing the way we study economic, agricultural and demographic trends world-wide. And the global Internet offers an extraordinary new tool for behavioural research. Epidemiologists have studied the frequency with which people search online for keywords such as 'flu', as a way to monitor disease spread. Other researchers, trying to understand how people would react to pandemic alerts, have looked at the way online gamers in 'World of Warcraft' congregate around the digital equivalent of disaster zones, as a clue to new disease-control strategies.[11]



Evolution of ESA's EO Data Archives between 1986-2007 and future estimates (up to 2020)

Total Archive in Terrabytes (TB)
22000
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1986
1989
1993
1995
1998
2000
2003
2005
2007
2015
2020
Future Date Estimates
LANDSAT 2-4 MSS (75 - Dec 93)
AQUA Modis (April 03 - today)
ENVISAT LR (March 02 - today)
ENVISAT HR (March 02 - today)
TERRA Modis (June 01 - today)
QUICK SCATT (01 - today / PROBA (May 02 - today)
LANDSAT 7 ETM (April 99 - Dec 03)
SEA STAR SeaWifs (April 98 - today)
ERS 2 HR (May 95 - today)
ERS 2 LBR (May 95 - today)
JERS SAR/OPS VNIR (92 - Sep 98)
ERS 1 HR (Jul 91 - Mar 00)
ERS 1 LBR (Jul 91 - Mar 00)
SPOT 1-4 HRV (87 - today)
MOS 1, 1b MESSR (87 - today)
NOAA 9-17 AVHRR (86 - today)
LANDSAT 5 TM (April 84 - today)
NIMBUS 7 (Nov 78 - May 86), SEASAT (Jun-Oct 78)

Yearly Data Creation on NICE

-350
-300
-250
-200
-150
-100
-50
0
TBytes (1TB = 10^{12} Bytes)
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007

Left: After a long, slow rise since 1986, the volume of earth-observation data from the European Space Agency's satellites passed three petabytes in 2007 – three million, billion bytes. The projection for 2020: A seven-fold rise.

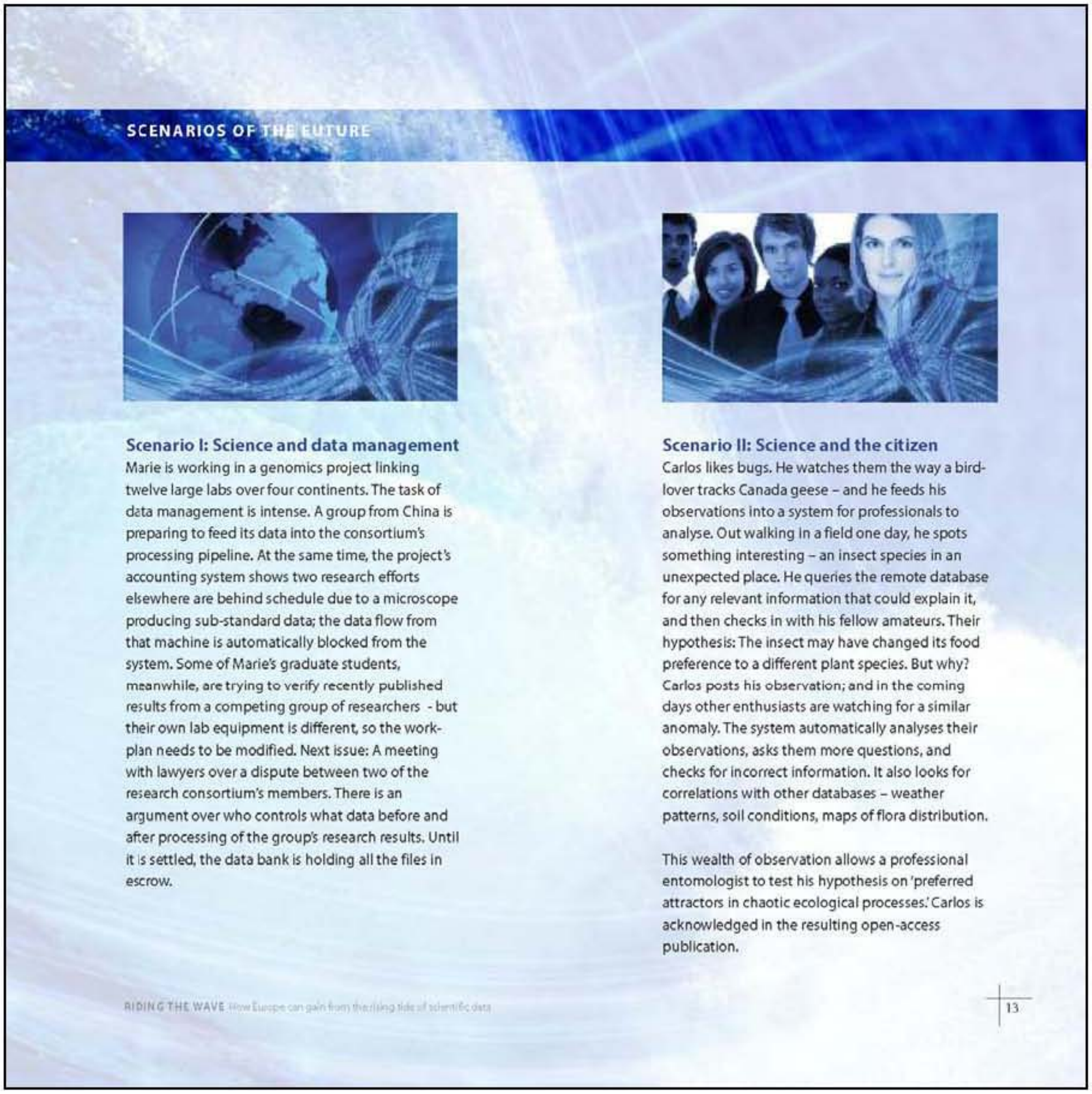
But it is not just what is studied, but also who studies it and how, that is affected by the data tide. For the first time in two centuries of growing professionalisation in science, new ICT tools allow the increasing involvement of amateurs; there are simply too many observations required for the professionals to do it all on their own. An example is GalaxyZoo, a Web portal through which amateurs help astronomers explore the Universe.[12] Biodiversity monitoring depends to a large extent on the power of observations by tens of thousands of volunteers who send their notes on species in defined areas to a central data repository.[13] For instance, the Swedish Species Gateway provides a navigable visual interface linked to geographical information systems.[14] These examples represent an important social and political trend. Empowering informed 'citizen-scientists' will also empower science.

Where will this lead? Consider five diverse scenarios that we believe to be entirely possible in coming decades.

Right: In Grenoble, the European Synchrotron Radiation Facility is a super-microscope studying anything from the propagation of cracks in steel to the surface proteins on the influenza virus. In the decade to 2007, its annual data output rose more than a hundred-fold. And it is just one of about 50 synchrotrons world-wide.

12

RIDING THE WAVE How Europe can gain from the rising tide of scientific data



SCENARIOS OF THE FUTURE

Scenario I: Science and data management

Marie is working in a genomics project linking twelve large labs over four continents. The task of data management is intense. A group from China is preparing to feed its data into the consortium's processing pipeline. At the same time, the project's accounting system shows two research efforts elsewhere are behind schedule due to a microscope producing sub-standard data; the data flow from that machine is automatically blocked from the system. Some of Marie's graduate students, meanwhile, are trying to verify recently published results from a competing group of researchers - but their own lab equipment is different, so the work-plan needs to be modified. Next issue: A meeting with lawyers over a dispute between two of the research consortium's members. There is an argument over who controls what data before and after processing of the group's research results. Until it is settled, the data bank is holding all the files in escrow.

Scenario II: Science and the citizen

Carlos likes bugs. He watches them the way a bird-lover tracks Canada geese – and he feeds his observations into a system for professionals to analyse. Out walking in a field one day, he spots something interesting – an insect species in an unexpected place. He queries the remote database for any relevant information that could explain it, and then checks in with his fellow amateurs. Their hypothesis: The insect may have changed its food preference to a different plant species. But why? Carlos posts his observation; and in the coming days other enthusiasts are watching for a similar anomaly. The system automatically analyses their observations, asks them more questions, and checks for incorrect information. It also looks for correlations with other databases – weather patterns, soil conditions, maps of flora distribution.

This wealth of observation allows a professional entomologist to test his hypothesis on 'preferred attractors in chaotic ecological processes.' Carlos is acknowledged in the resulting open-access publication.

RIDING THE WAVE How Europe can gain from the rising tide of scientific data 13

SCENARIOS OF THE FUTURE

Scenario III: Science and the data set

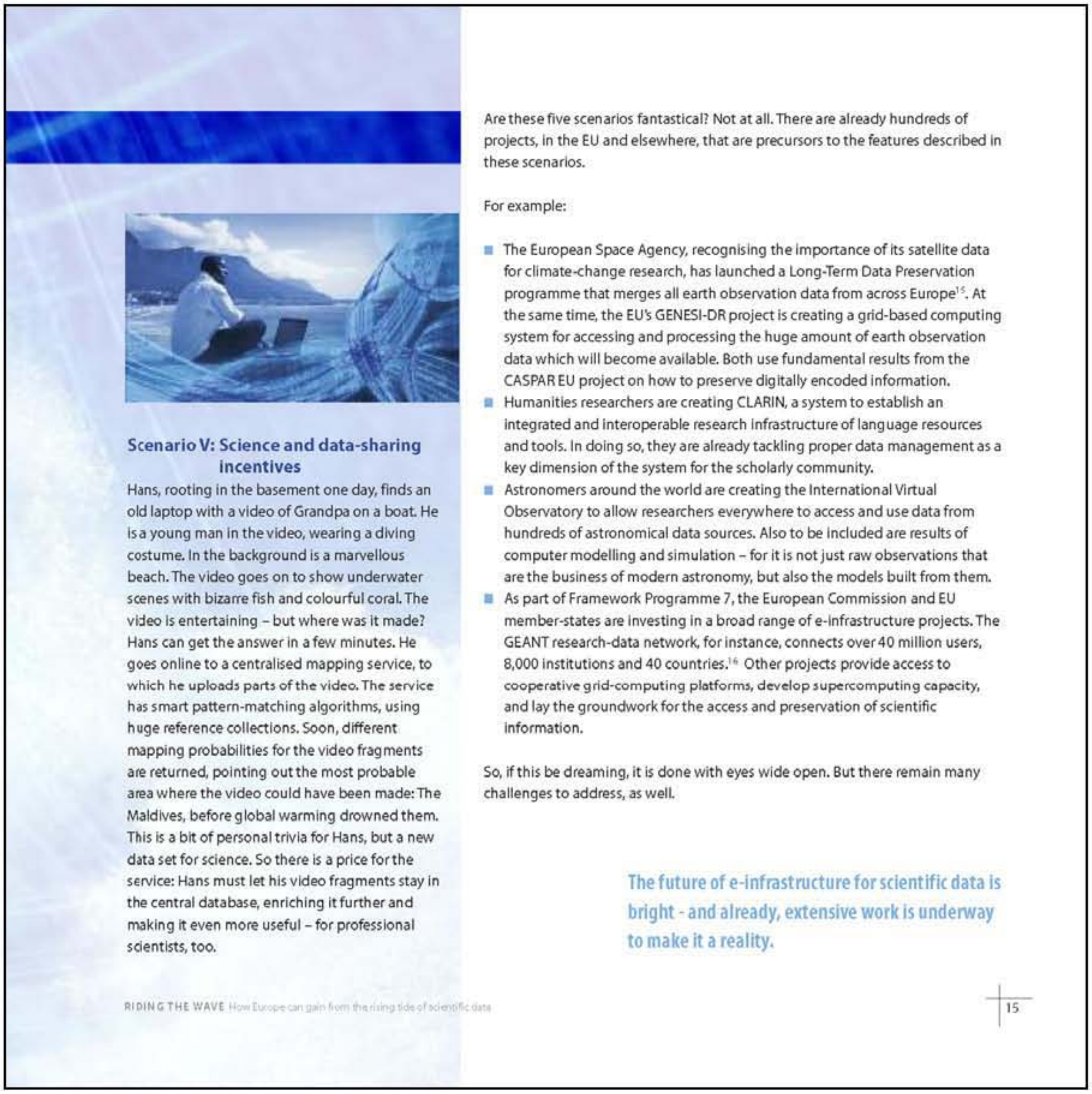
Anneli has a grant under which she is allowed free access to 10 years of measurements by the global cell-phone sensor network, stored in cross-continental archives. This network uses the miniature sensors standard in cell phones to monitor local temperature, air quality, wind speed, light intensity, noise levels and other parameters – and links it to GPS information. The information is all kept in regional archives, with open interfaces so researchers can query them uniformly. With her team, Anneli wants to investigate correlations between the environment and the spread of illness – and for the disease information, she is looking at anonymised, geo-tagged messages sent by people mentioning the disease. She intends to clean the resulting data set and make it publicly available via her university's institutional repository. From there, it could become the scientific equivalent of a Top-40 song – played by others around the world. Her chances for tenure rise.

Scenario IV: Science and the student

Roger is working on an international PhD. It's a relatively new programme, in which a student applies to become a member of an international team working on a big problem that affects all people. His group is comparing many forms of non-verbal communications between cultures. It has several hundred members and his university tutor is one of the nodal points contributing expertise in 'synergistic communication between biological components.' Others in the network are using archaeological evidence to study communications between ancient Mesopotamian and Hellenic cultures; some are studying computer-computer interactions between different systems; yet more are studying communications in refugee camps. Each node contributes to the whole. Results are communicated as they happen, and there are daily, virtual-presence planning sessions. Roger had to sign a contract not to misuse data or contribute anything that is not for the common good – such as externally sourced information that he has not thoroughly checked for provenance.

14

RIDING THE WAVE How Europe can gain from the rising tide of scientific data



Scenario V: Science and data-sharing incentives

Hans, rooting in the basement one day, finds an old laptop with a video of Grandpa on a boat. He is a young man in the video, wearing a diving costume. In the background is a marvellous beach. The video goes on to show underwater scenes with bizarre fish and colourful coral. The video is entertaining – but where was it made? Hans can get the answer in a few minutes. He goes online to a centralised mapping service, to which he uploads parts of the video. The service has smart pattern-matching algorithms, using huge reference collections. Soon, different mapping probabilities for the video fragments are returned, pointing out the most probable area where the video could have been made: The Maldives, before global warming drowned them. This is a bit of personal trivia for Hans, but a new data set for science. So there is a price for the service: Hans must let his video fragments stay in the central database, enriching it further and making it even more useful – for professional scientists, too.

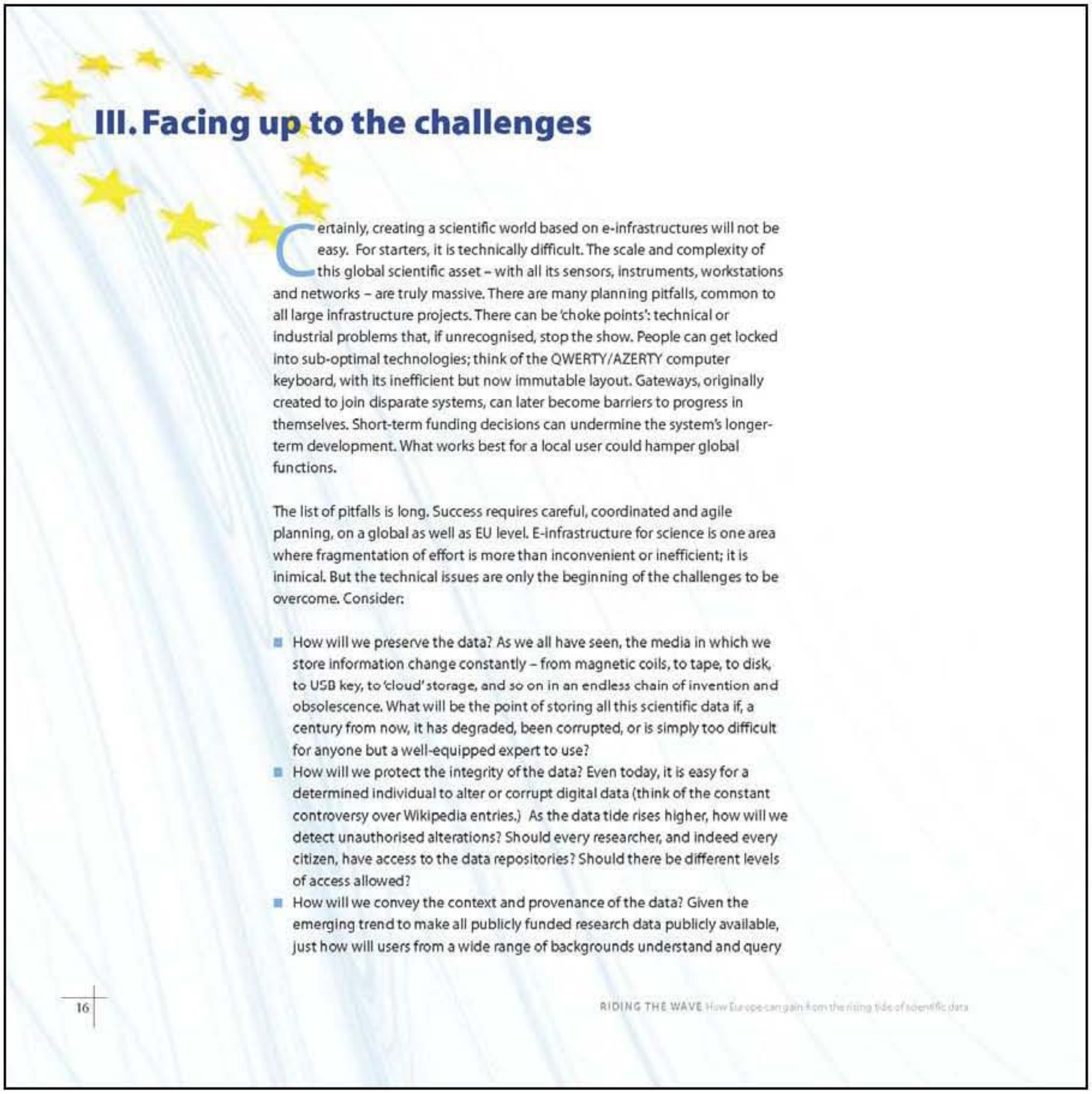
Are these five scenarios fantastical? Not at all. There are already hundreds of projects, in the EU and elsewhere, that are precursors to the features described in these scenarios.

For example:

- The European Space Agency, recognising the importance of its satellite data for climate-change research, has launched a Long-Term Data Preservation programme that merges all earth observation data from across Europe[15]. At the same time, the EU's GENESI-DR project is creating a grid-based computing system for accessing and processing the huge amount of earth observation data which will become available. Both use fundamental results from the CASPAR EU project on how to preserve digitally encoded information.
- Humanities researchers are creating CLARIN, a system to establish an integrated and interoperable research infrastructure of language resources and tools. In doing so, they are already tackling proper data management as a key dimension of the system for the scholarly community.
- Astronomers around the world are creating the International Virtual Observatory to allow researchers everywhere to access and use data from hundreds of astronomical data sources. Also to be included are results of computer modelling and simulation – for it is not just raw observations that are the business of modern astronomy, but also the models built from them.
- As part of Framework Programme 7, the European Commission and EU member-states are investing in a broad range of e-infrastructure projects. The GEANT research-data network, for instance, connects over 40 million users, 8,000 institutions and 40 countries.[16] Other projects provide access to cooperative grid-computing platforms, develop supercomputing capacity, and lay the groundwork for the access and preservation of scientific information.

So, if this be dreaming, it is done with eyes wide open. But there remain many challenges to address, as well.

The future of e-infrastructure for scientific data is bright - and already, extensive work is underway to make it a reality.

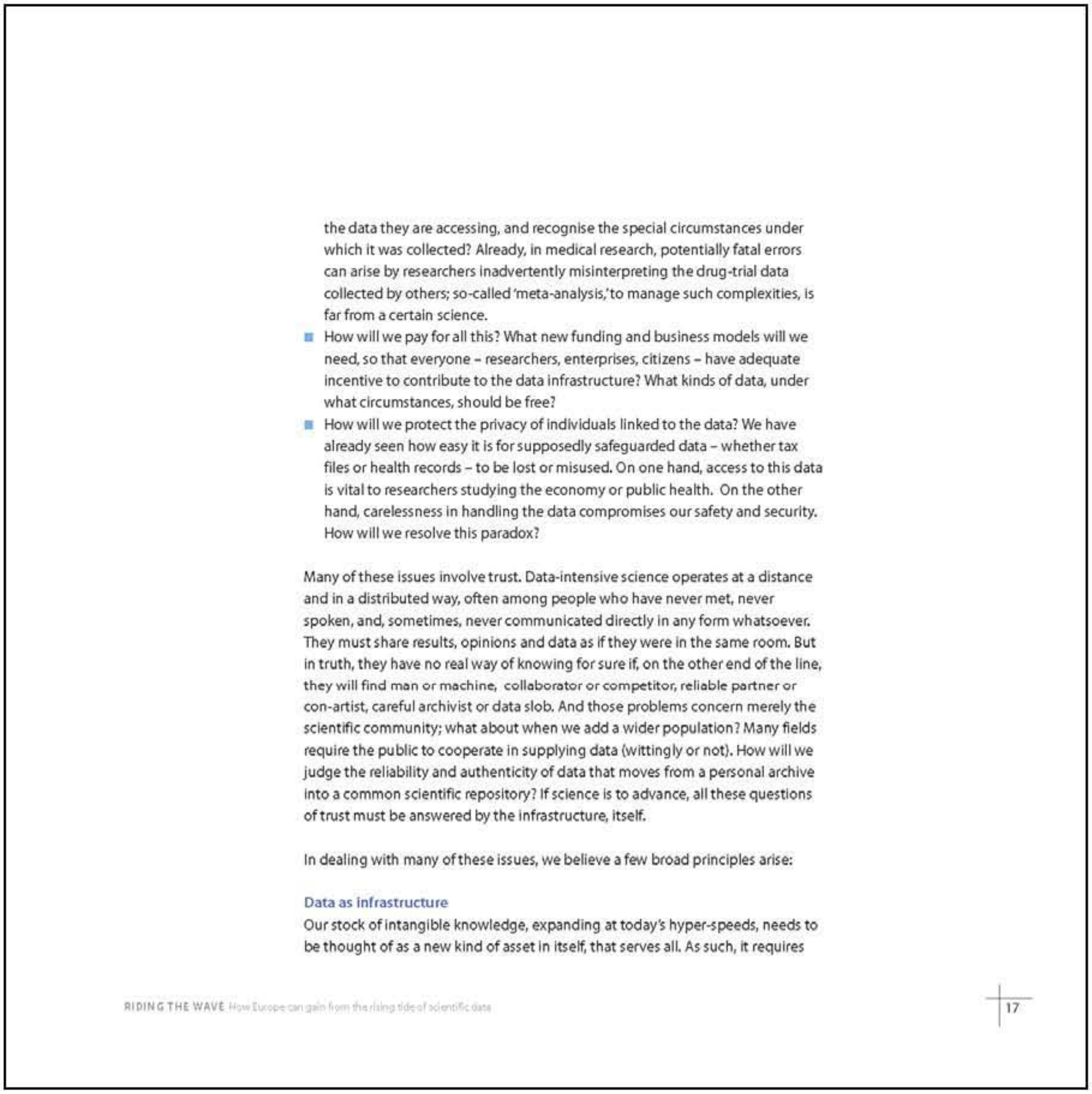


III. Facing up to the challenges

Certainly, creating a scientific world based on e-infrastructures will not be easy. For starters, it is technically difficult. The scale and complexity of this global scientific asset – with all its sensors, instruments, workstations and networks – are truly massive. There are many planning pitfalls, common to all large infrastructure projects. There can be 'choke points': technical or industrial problems that, if unrecognised, stop the show. People can get locked into sub-optimal technologies; think of the QWERTY/AZERTY computer keyboard, with its inefficient but now immutable layout. Gateways, originally created to join disparate systems, can later become barriers to progress in themselves. Short-term funding decisions can undermine the system's longer-term development. What works best for a local user could hamper global functions.

The list of pitfalls is long. Success requires careful, coordinated and agile planning, on a global as well as EU level. E-infrastructure for science is one area where fragmentation of effort is more than inconvenient or inefficient; it is inimical. But the technical issues are only the beginning of the challenges to be overcome. Consider:

- How will we preserve the data? As we all have seen, the media in which we store information change constantly – from magnetic coils, to tape, to disk, to USB key, to 'cloud' storage, and so on in an endless chain of invention and obsolescence. What will be the point of storing all this scientific data if, a century from now, it has degraded, been corrupted, or is simply too difficult for anyone but a well-equipped expert to use?
- How will we protect the integrity of the data? Even today, it is easy for a determined individual to alter or corrupt digital data (think of the constant controversy over Wikipedia entries.) As the data tide rises higher, how will we detect unauthorised alterations? Should every researcher, and indeed every citizen, have access to the data repositories? Should there be different levels of access allowed?
- How will we convey the context and provenance of the data? Given the emerging trend to make all publicly funded research data publicly available, just how will users from a wide range of backgrounds understand and query



the data they are accessing, and recognise the special circumstances under which it was collected? Already, in medical research, potentially fatal errors can arise by researchers inadvertently misinterpreting the drug-trial data collected by others; so-called 'meta-analysis,' to manage such complexities, is far from a certain science.

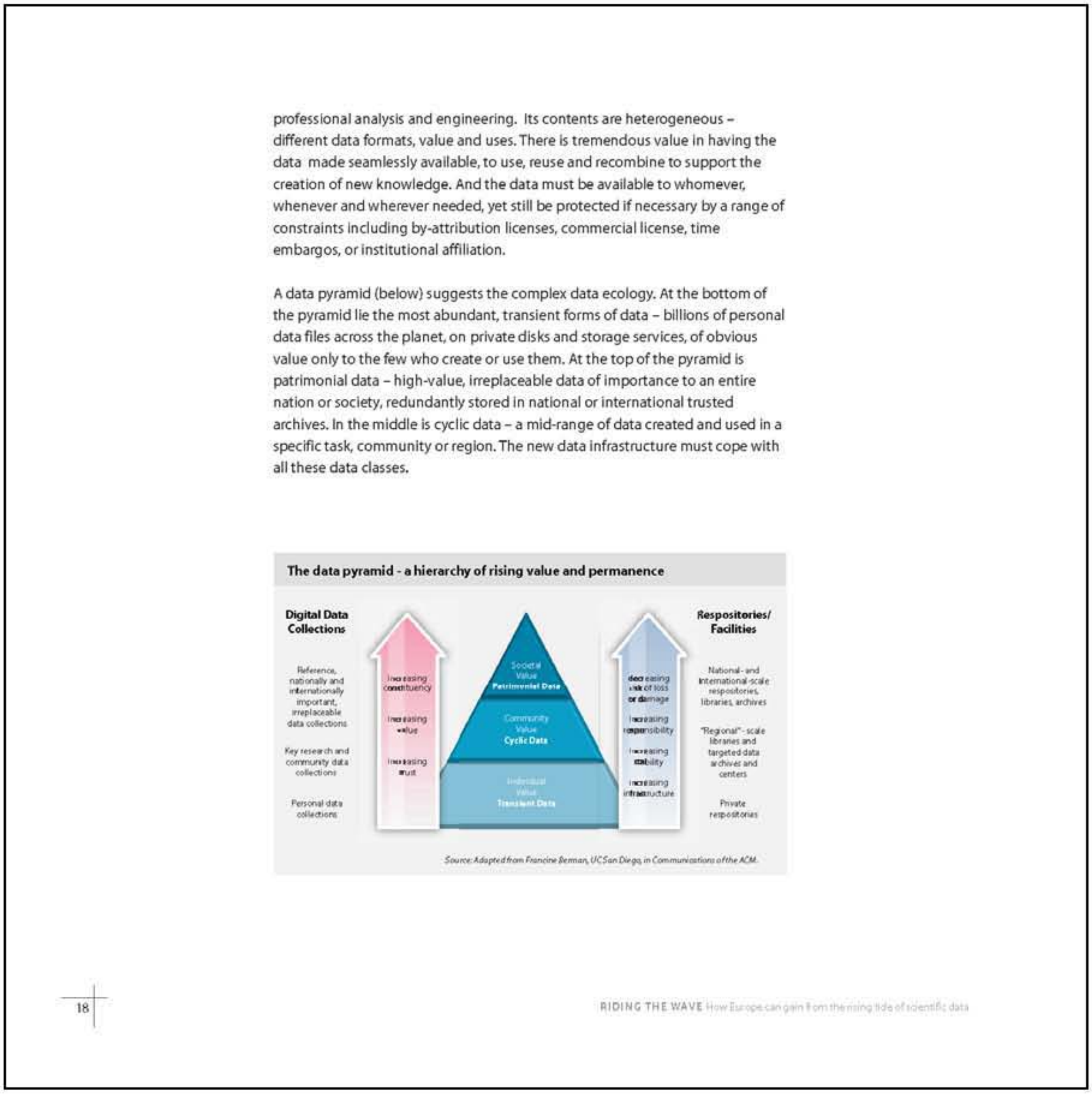
- How will we pay for all this? What new funding and business models will we need, so that everyone – researchers, enterprises, citizens – have adequate incentive to contribute to the data infrastructure? What kinds of data, under what circumstances, should be free?
- How will we protect the privacy of individuals linked to the data? We have already seen how easy it is for supposedly safeguarded data – whether tax files or health records – to be lost or misused. On one hand, access to this data is vital to researchers studying the economy or public health. On the other hand, carelessness in handling the data compromises our safety and security. How will we resolve this paradox?

Many of these issues involve trust. Data-intensive science operates at a distance and in a distributed way, often among people who have never met, never spoken, and, sometimes, never communicated directly in any form whatsoever. They must share results, opinions and data as if they were in the same room. But in truth, they have no real way of knowing for sure if, on the other end of the line, they will find man or machine, collaborator or competitor, reliable partner or con-artist, careful archivist or data slob. And those problems concern merely the scientific community; what about when we add a wider population? Many fields require the public to cooperate in supplying data (wittingly or not). How will we judge the reliability and authenticity of data that moves from a personal archive into a common scientific repository? If science is to advance, all these questions of trust must be answered by the infrastructure, itself.

In dealing with many of these issues, we believe a few broad principles arise:

Data as infrastructure

Our stock of intangible knowledge, expanding at today's hyper-speeds, needs to be thought of as a new kind of asset in itself, that serves all. As such, it requires



professional analysis and engineering. Its contents are heterogeneous – different data formats, value and uses. There is tremendous value in having the data made seamlessly available, to use, reuse and recombine to support the creation of new knowledge. And the data must be available to whomever, whenever and wherever needed, yet still be protected if necessary by a range of constraints including by-attribution licenses, commercial license, time embargos, or institutional affiliation.

A data pyramid (below) suggests the complex data ecology. At the bottom of the pyramid lie the most abundant, transient forms of data – billions of personal data files across the planet, on private disks and storage services, of obvious value only to the few who create or use them. At the top of the pyramid is patrimonial data – high-value, irreplaceable data of importance to an entire nation or society, redundantly stored in national or international trusted archives. In the middle is cyclic data – a mid-range of data created and used in a specific task, community or region. The new data infrastructure must cope with all these data classes.

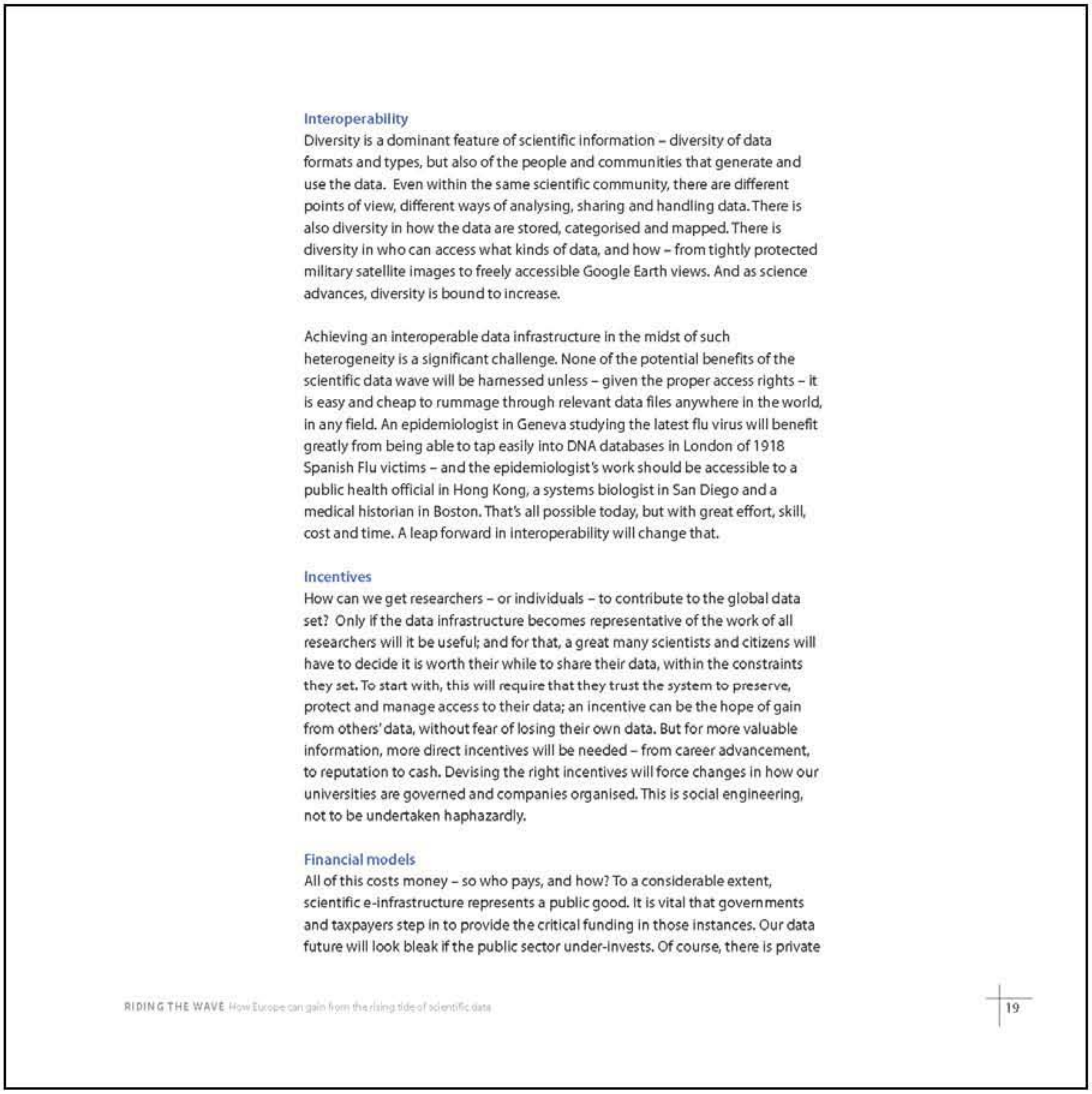
The data pyramid - a hierarchy of rising value and permanence

Digital Data Collections
Reference, nationally and internationally important, irreplaceable data collections
Key research and community data collections
Personal data collections
Increasing constituency
Increasing value
Increasing trust
Societal Value
Patrimonial Data
Community Value
Cyclic Data
Individual Value
Transient Data
decreasing risk of loss or damage
Increasing responsibility
Increasing stability
increasing infrastructure
Respositories/ Facilities
National- and International-scale respositories, libraries, archives
"Regional" - scale libraries and targeted data archives and centers
Private respositories

Source: Adapted from Francine Berman, UC San Diego, in Communications of the ACM.

18

RIDING THE WAVE How Europe can gain from the rising tide of scientific data



Interoperability

Diversity is a dominant feature of scientific information – diversity of data formats and types, but also of the people and communities that generate and use the data. Even within the same scientific community, there are different points of view, different ways of analysing, sharing and handling data. There is also diversity in how the data are stored, categorised and mapped. There is diversity in who can access what kinds of data, and how – from tightly protected military satellite images to freely accessible Google Earth views. And as science advances, diversity is bound to increase.

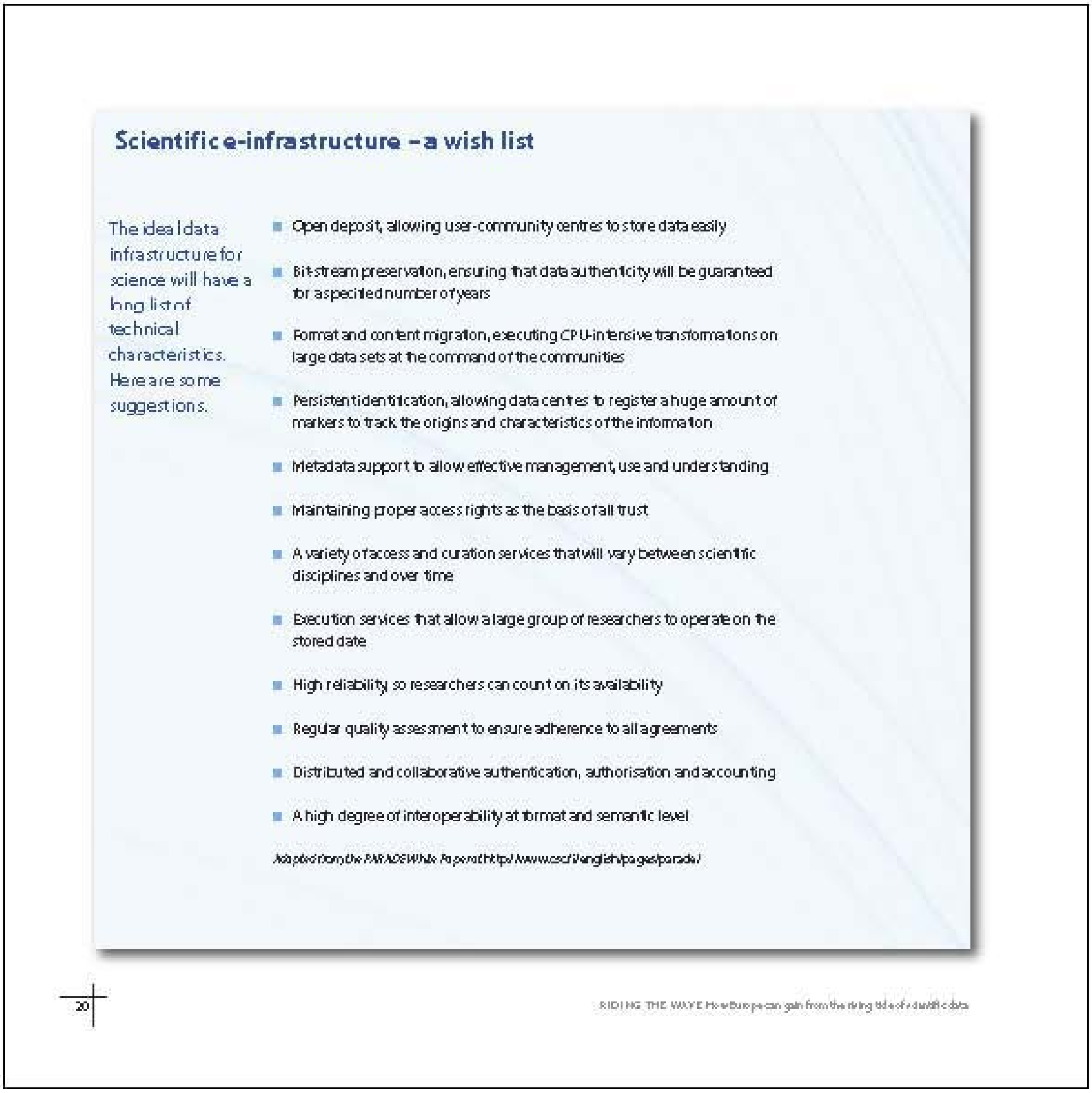
Achieving an interoperable data infrastructure in the midst of such heterogeneity is a significant challenge. None of the potential benefits of the scientific data wave will be harnessed unless – given the proper access rights – it is easy and cheap to rummage through relevant data files anywhere in the world, in any field. An epidemiologist in Geneva studying the latest flu virus will benefit greatly from being able to tap easily into DNA databases in London of 1918 Spanish Flu victims – and the epidemiologist's work should be accessible to a public health official in Hong Kong, a systems biologist in San Diego and a medical historian in Boston. That's all possible today, but with great effort, skill, cost and time. A leap forward in interoperability will change that.

Incentives

How can we get researchers – or individuals – to contribute to the global data set? Only if the data infrastructure becomes representative of the work of all researchers will it be useful; and for that, a great many scientists and citizens will have to decide it is worth their while to share their data, within the constraints they set. To start with, this will require that they trust the system to preserve, protect and manage access to their data; an incentive can be the hope of gain from others' data, without fear of losing their own data. But for more valuable information, more direct incentives will be needed – from career advancement, to reputation to cash. Devising the right incentives will force changes in how our universities are governed and companies organised. This is social engineering, not to be undertaken haphazardly.

Financial models

All of this costs money – so who pays, and how? To a considerable extent, scientific e-infrastructure represents a public good. It is vital that governments and taxpayers step in to provide the critical funding in those instances. Our data future will look bleak if the public sector under-invests. Of course, there is private



Scientific e-infrastructure – a wish list

The ideal data infrastructure for science will have a long list of technical characteristics. Here are some suggestions.

- Open deposit, allowing user-community centres to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities
- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored date
- High reliability so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorisation and accounting
- A high degree of interoperability at format and semantic level

Adapted from the PARADE White Paper at http://www.csc.fi/english/pages/parade/

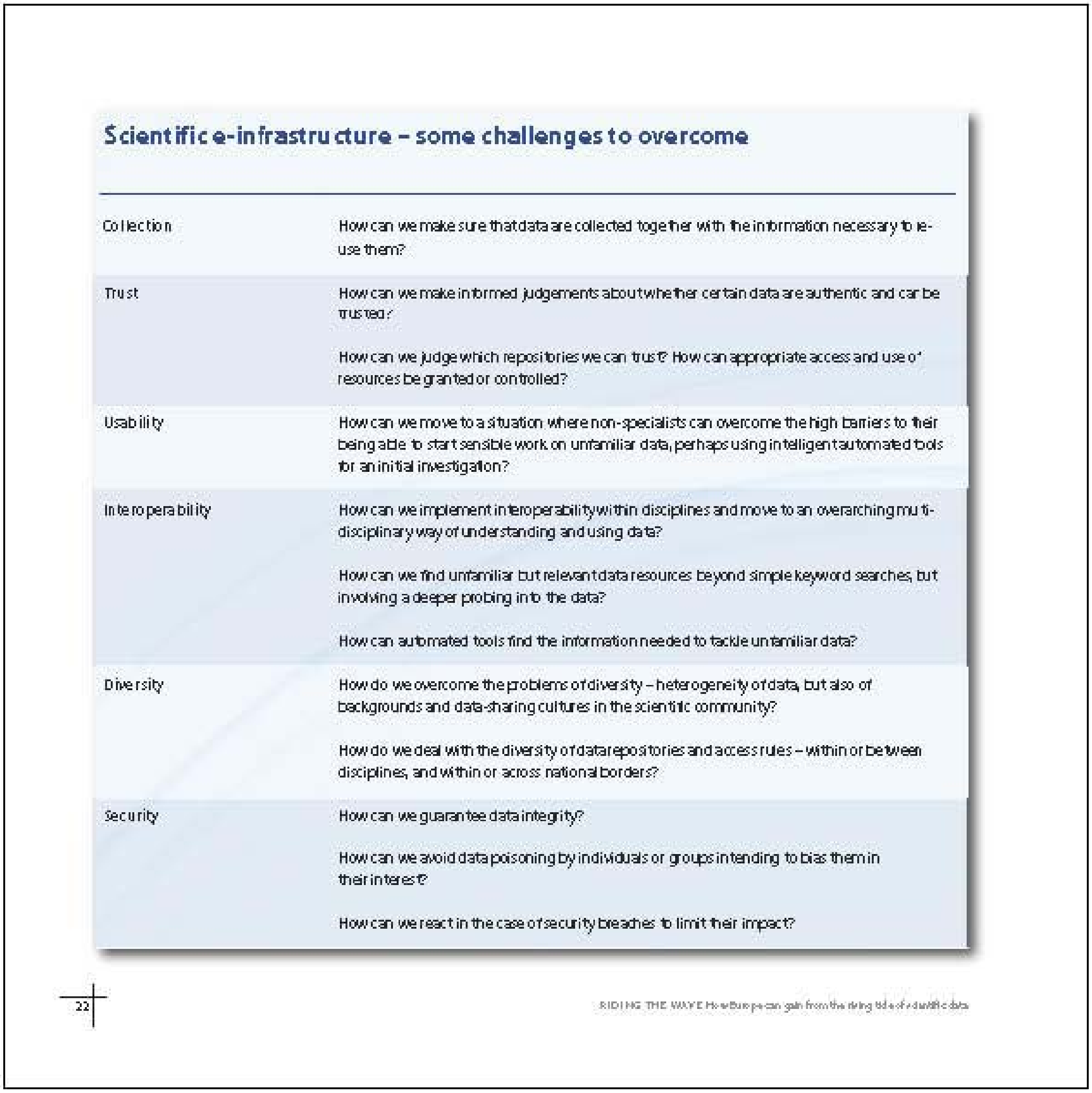
20

RIDING THE WAVE How Europe can gain from the rising tide of scientific data

gain as well. When a government laboratory contributes its raw research data to the global e-infrastructure, it is certainly saving private users the expense of running those experiments on their own. Equally, when a private company contributes its own files to the system, it also helps the public researchers. It is important to devise funding mechanisms that enable all to contribute as well as to benefit, through an increased return on investment.

These issues can be resolved. We have experience of past changes in how we store, share and manage valuable assets. As the technology of food and transport evolved, society moved from self-supporting farmers to town markets, and from markets to a range of supermarkets and specialty shops. In finance, we moved from private hoards to communal banks to international markets. The same path from individual control to international exchange must be trodden by data – indeed, it is already happening.

It is important to devise funding mechanisms that enable all to contribute as well as to benefit, through an increased return on investment.

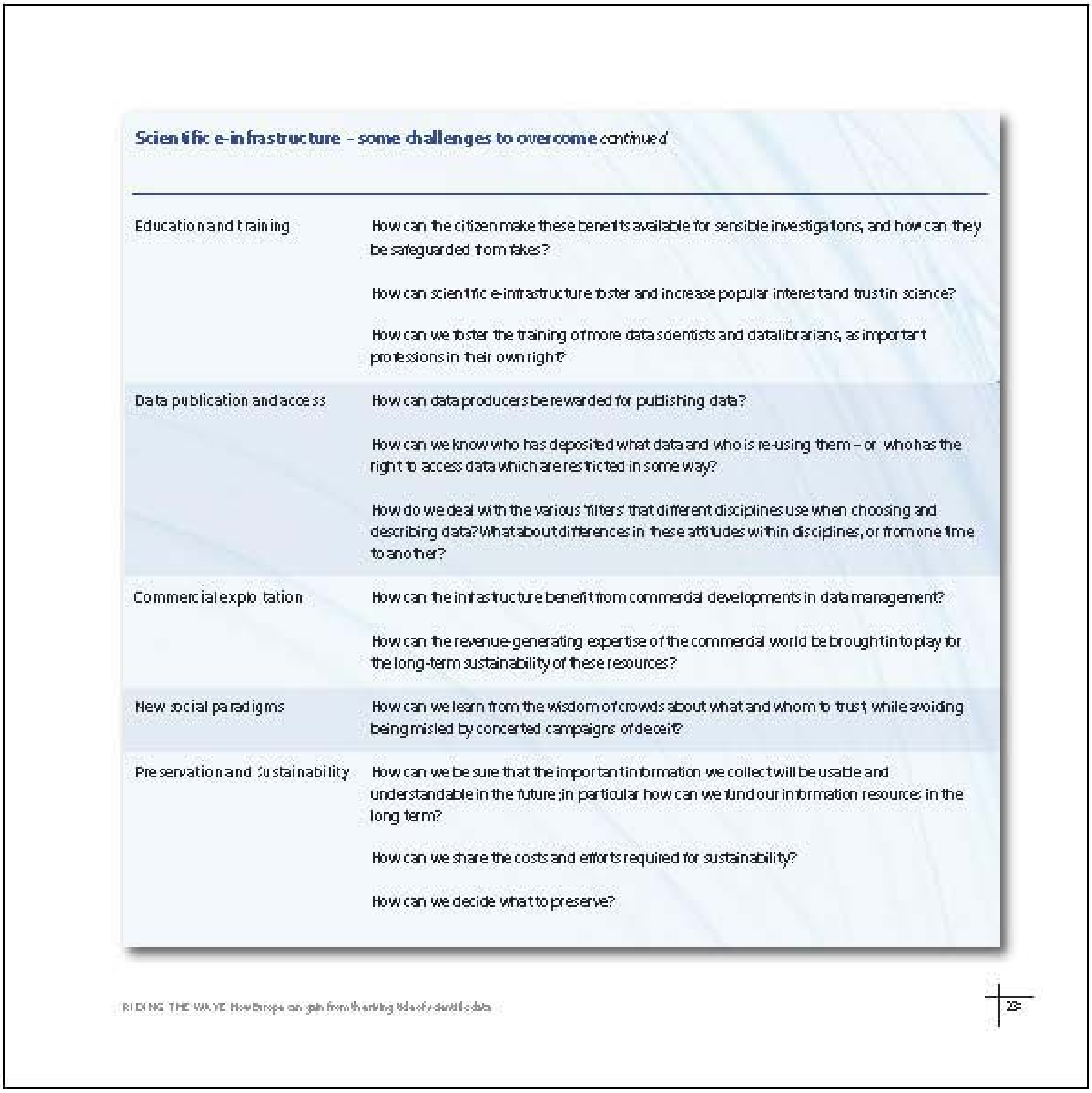


Scientific e-infrastructure – some challenges to overcome

Collection	How can we make sure that data are collected together with the information necessary to re-use them?
Trust	How can we make informed judgements about whether certain data are authentic and can be trusted? How can we judge which repositories we can trust? How can appropriate access and use of resources be granted or controlled?
Usability	How can we move to a situation where non-specialists can overcome the high barriers to their being able to start sensible work on unfamiliar data, perhaps using intelligent automated tools for an initial investigation?
Interoperability	How can we implement interoperability within disciplines and move to an overarching multi-disciplinary way of understanding and using data? How can we find unfamiliar but relevant data resources beyond simple keyword searches, but involving a deeper probing into the data? How can automated tools find the information needed to tackle unfamiliar data?
Diversity	How do we overcome the problems of diversity – heterogeneity of data, but also of backgrounds and data-sharing cultures in the scientific community? How do we deal with the diversity of data repositories and access rules – within or between disciplines, and within or across national borders?
Security	How can we guarantee data integrity? How can we avoid data poisoning by individuals or groups intending to bias them in their interest? How can we react in the case of security breaches to limit their impact?

22

RIDING THE WAVE How Europe can gain from the rising tide of scientific data



Scientific e-infrastructure – some challenges to overcome *continued*

Education and training	How can the citizen make these benefits available for sensible investigations, and how can they be safeguarded from fakes? How can scientific e-infrastructure foster and increase popular interest and trust in science? How can we foster the training of more data scientists and data librarians, as important professions in their own right?
Data publication and access	How can data producers be rewarded for publishing data? How can we know who has deposited what data and who is re-using them – or who has the right to access data which are restricted in some way? How do we deal with the various 'filters' that different disciplines use when choosing and describing data? What about differences in these attitudes within disciplines, or from one time to another?
Commercial exploitation	How can the infrastructure benefit from commercial developments in data management? How can the revenue-generating expertise of the commercial world be brought into play for the long-term sustainability of these resources?
New social paradigms	How can we learn from the wisdom of crowds about what and whom to trust, while avoiding being misled by concerted campaigns of deceit?
Preservation and sustainability	How can we be sure that the important information we collect will be usable and understandable in the future; in particular how can we fund our information resources in the long term? How can we share the costs and efforts required for sustainability? How can we decide what to preserve?

RIDING THE WAVE How Europe can gain from the rising tide of scientific data 23

IV. A vision for 2030

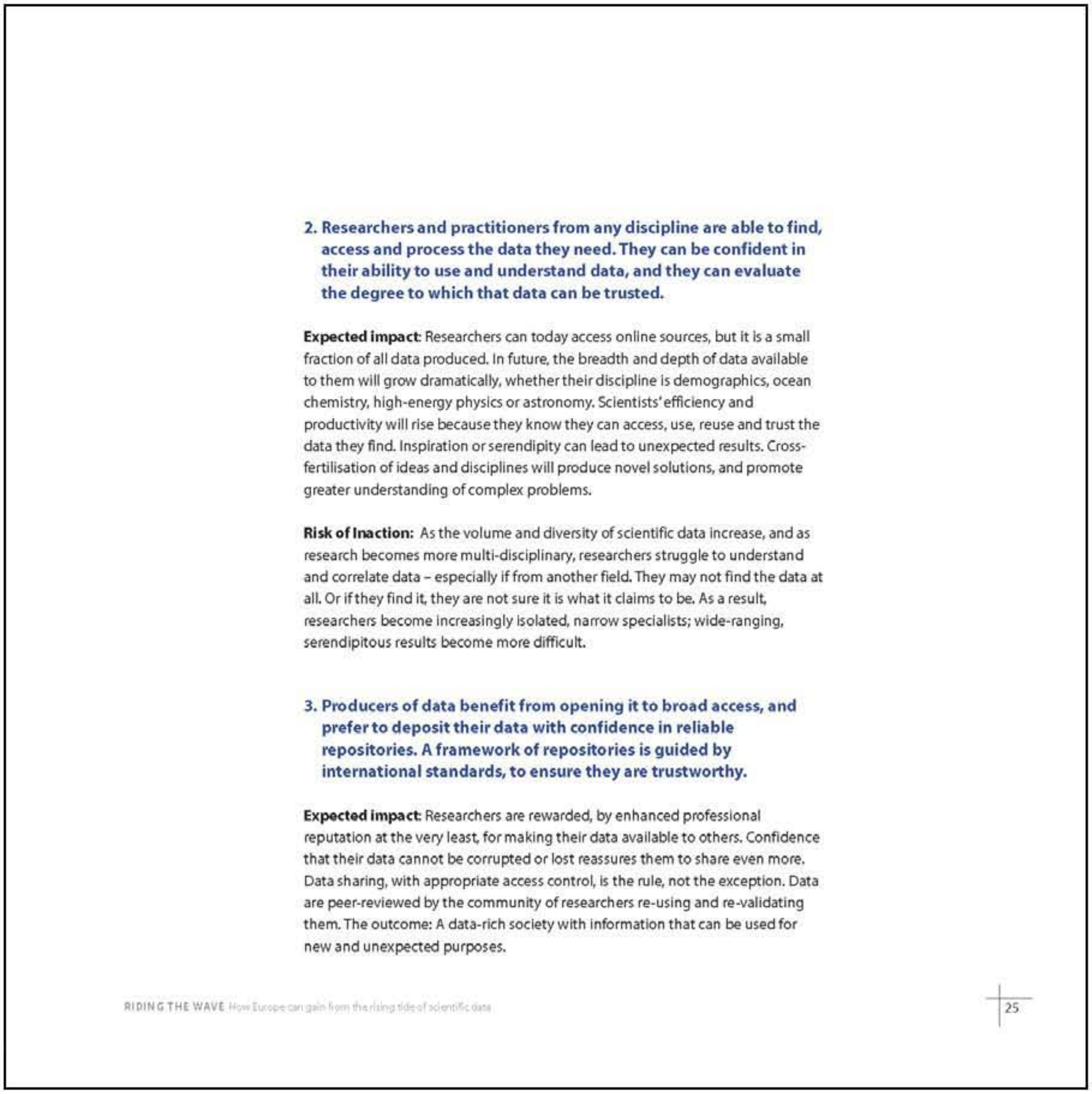
The creation of scientific e-infrastructure is a means, not an end. It is a means to new science, new solutions and new progress in society. We cannot predict what the world will be like in 2030, but we can state some broad principles of what it should be like if scientific e-infrastructure is by then the major contributor to society, the economy and science that we expect it to be. All of these principles – our vision - point in the direction of an infrastructure that supports seamless access, use, reuse, and trust of data. It suggests a future in which the data infrastructure becomes invisible, and the data themselves have become infrastructure – a valuable asset, on which science, technology, the economy and society can advance. We will know we are well on our way to realising this vision when we see the following milestones achieved:

1. All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.

This may sound obvious – but it is by no means so.[17] Today we see the relative priorities of society in constant flux on the Internet and other electronic media. In a world of limited resources, how urgent is a packet of scientific data compared to home videos? How much is it worth to create reliable back-up and storage systems for what may seem today like transient chat messages, but could tomorrow become vital behavioural or epidemiological data? Thus, the first task is simply to get the message out that scientific e-infrastructure is important to society.

Expected impact: The intellectual capital of Europe is used to generate economic and scientific advances now, and that capital is safely preserved for further exploitation by future generations.

Risk of Inaction: Resources for funding take a back seat to more pressing concerns, and data decays through neglect. When critical data – whether about climate, new medicines or historic monuments – are needed later on, it will be too late.



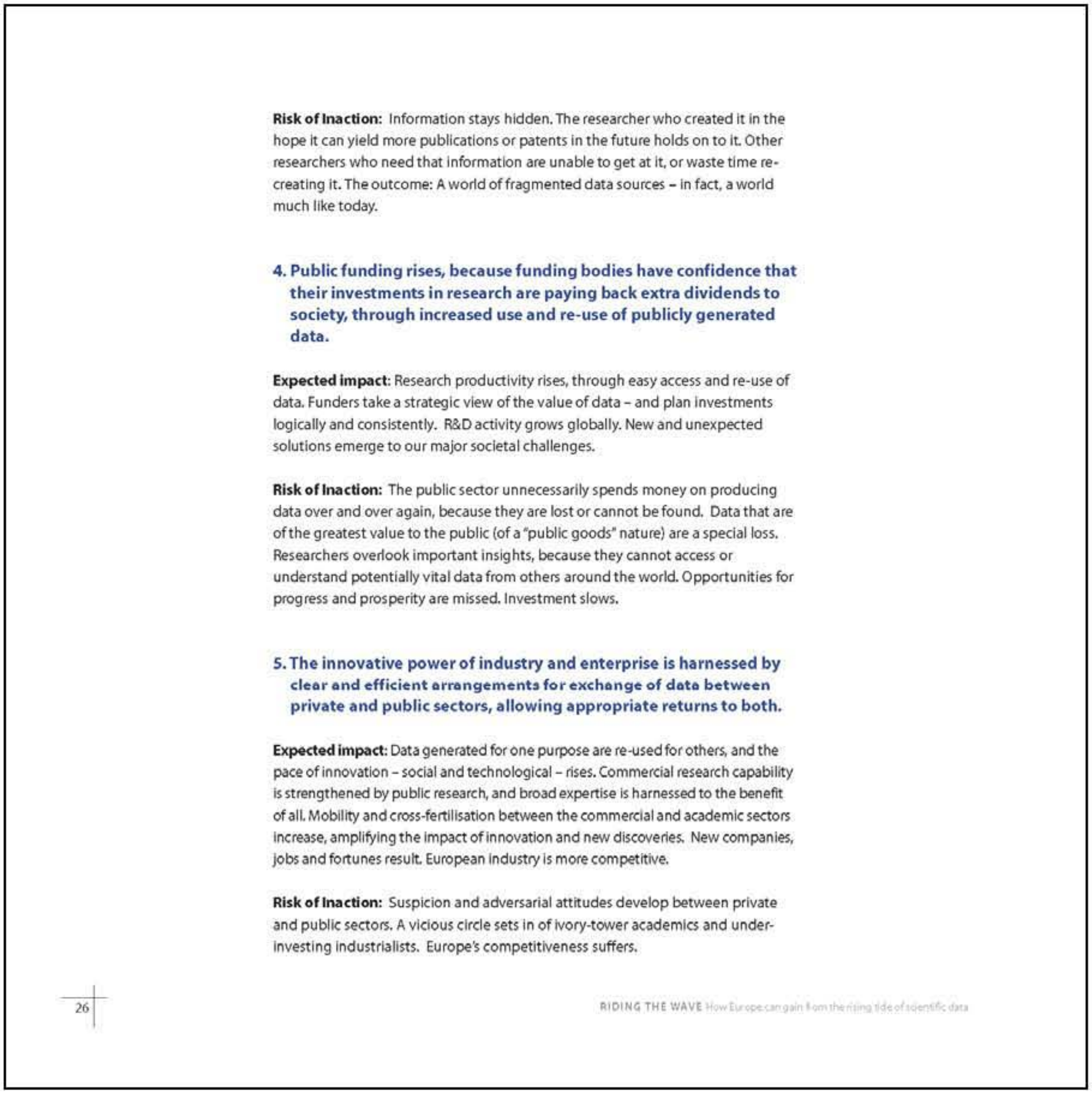
2. Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.

Expected impact: Researchers can today access online sources, but it is a small fraction of all data produced. In future, the breadth and depth of data available to them will grow dramatically, whether their discipline is demographics, ocean chemistry, high-energy physics or astronomy. Scientists' efficiency and productivity will rise because they know they can access, use, reuse and trust the data they find. Inspiration or serendipity can lead to unexpected results. Cross-fertilisation of ideas and disciplines will produce novel solutions, and promote greater understanding of complex problems.

Risk of Inaction: As the volume and diversity of scientific data increase, and as research becomes more multi-disciplinary, researchers struggle to understand and correlate data – especially if from another field. They may not find the data at all. Or if they find it, they are not sure it is what it claims to be. As a result, researchers become increasingly isolated, narrow specialists; wide-ranging, serendipitous results become more difficult.

3. Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy.

Expected impact: Researchers are rewarded, by enhanced professional reputation at the very least, for making their data available to others. Confidence that their data cannot be corrupted or lost reassures them to share even more. Data sharing, with appropriate access control, is the rule, not the exception. Data are peer-reviewed by the community of researchers re-using and re-validating them. The outcome: A data-rich society with information that can be used for new and unexpected purposes.



Risk of Inaction: Information stays hidden. The researcher who created it in the hope it can yield more publications or patents in the future holds on to it. Other researchers who need that information are unable to get at it, or waste time re-creating it. The outcome: A world of fragmented data sources – in fact, a world much like today.

4. Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.

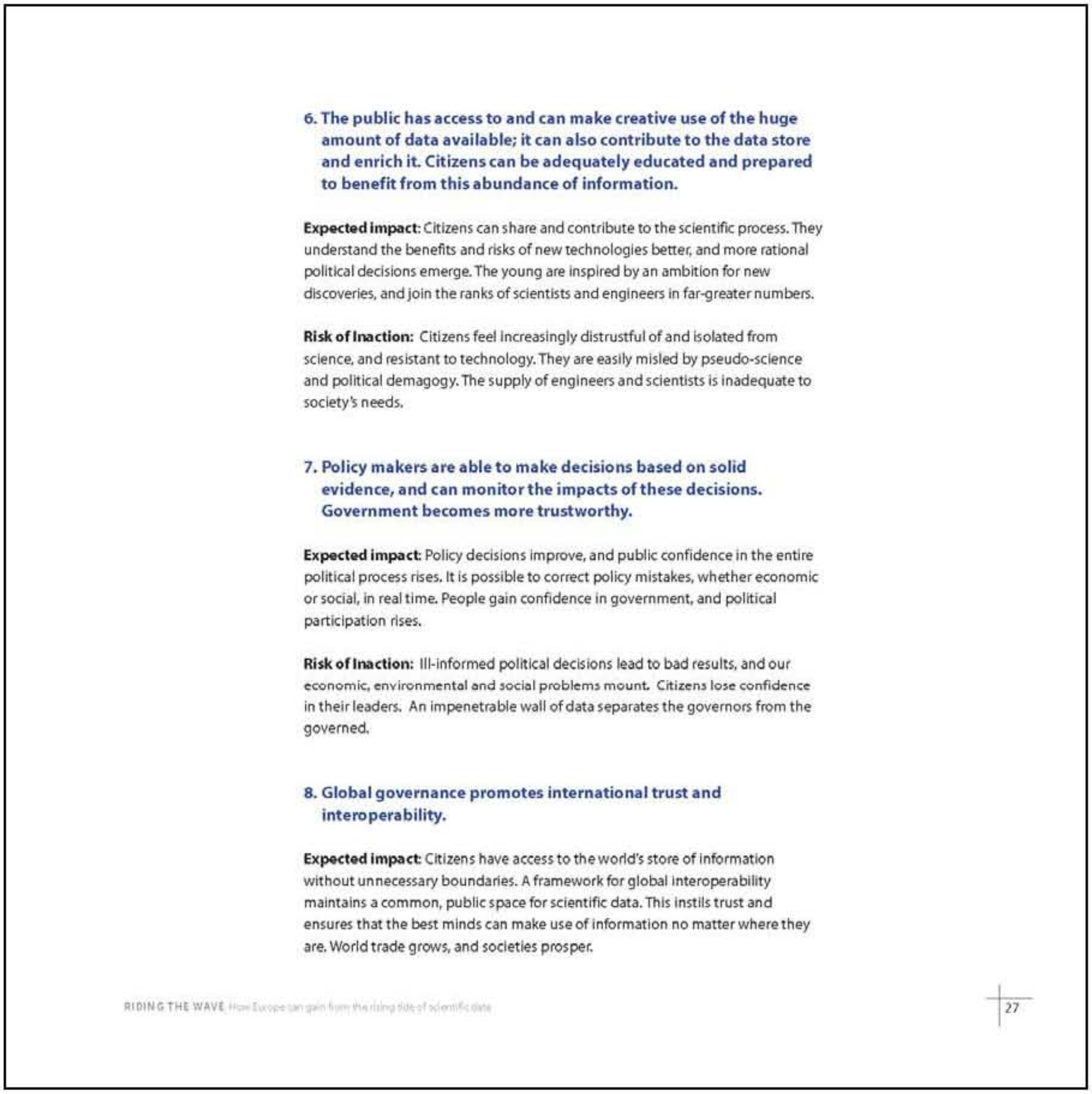
Expected impact: Research productivity rises, through easy access and re-use of data. Funders take a strategic view of the value of data – and plan investments logically and consistently. R&D activity grows globally. New and unexpected solutions emerge to our major societal challenges.

Risk of Inaction: The public sector unnecessarily spends money on producing data over and over again, because they are lost or cannot be found. Data that are of the greatest value to the public (of a "public goods" nature) are a special loss. Researchers overlook important insights, because they cannot access or understand potentially vital data from others around the world. Opportunities for progress and prosperity are missed. Investment slows.

5. The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.

Expected impact: Data generated for one purpose are re-used for others, and the pace of innovation – social and technological – rises. Commercial research capability is strengthened by public research, and broad expertise is harnessed to the benefit of all. Mobility and cross-fertilisation between the commercial and academic sectors increase, amplifying the impact of innovation and new discoveries. New companies, jobs and fortunes result. European industry is more competitive.

Risk of Inaction: Suspicion and adversarial attitudes develop between private and public sectors. A vicious circle sets in of ivory-tower academics and under-investing industrialists. Europe's competitiveness suffers.



6. The public has access to and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it. Citizens can be adequately educated and prepared to benefit from this abundance of information.

Expected impact: Citizens can share and contribute to the scientific process. They understand the benefits and risks of new technologies better, and more rational political decisions emerge. The young are inspired by an ambition for new discoveries, and join the ranks of scientists and engineers in far-greater numbers.

Risk of Inaction: Citizens feel increasingly distrustful of and isolated from science, and resistant to technology. They are easily misled by pseudo-science and political demagogy. The supply of engineers and scientists is inadequate to society's needs.

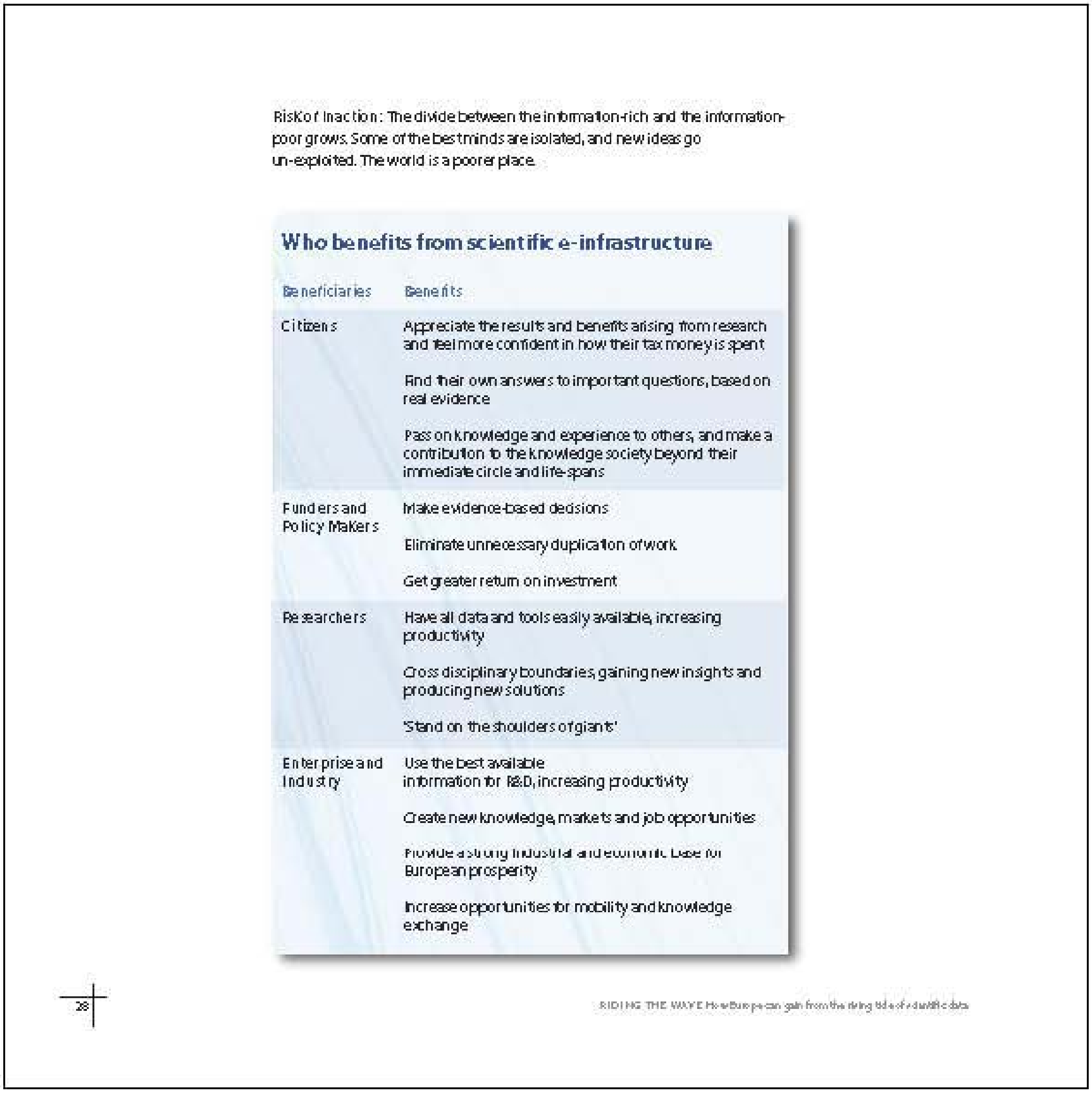
7. Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.

Expected impact: Policy decisions improve, and public confidence in the entire political process rises. It is possible to correct policy mistakes, whether economic or social, in real time. People gain confidence in government, and political participation rises.

Risk of Inaction: Ill-informed political decisions lead to bad results, and our economic, environmental and social problems mount. Citizens lose confidence in their leaders. An impenetrable wall of data separates the governors from the governed.

8. Global governance promotes international trust and interoperability.

Expected impact: Citizens have access to the world's store of information without unnecessary boundaries. A framework for global interoperability maintains a common, public space for scientific data. This instils trust and ensures that the best minds can make use of information no matter where they are. World trade grows, and societies prosper.



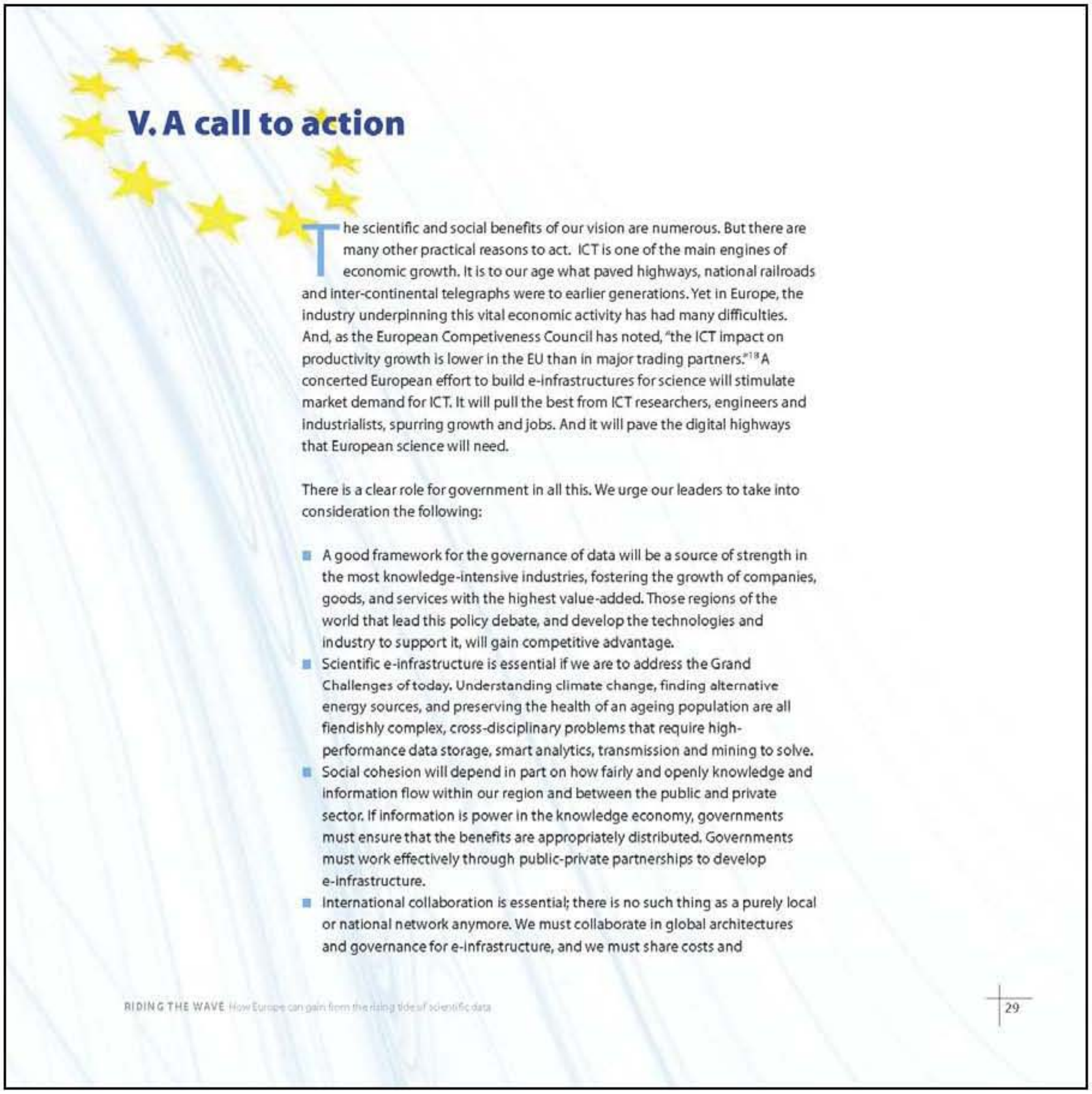
Risk of Inaction: The divide between the information-rich and the information-poor grows. Some of the best minds are isolated, and new ideas go un-exploited. The world is a poorer place.

Who benefits from scientific e-infrastructure

Beneficiaries	Benefits
Citizens	Appreciate the results and benefits arising from research and feel more confident in how their tax money is spent Find their own answers to important questions, based on real evidence Pass on knowledge and experience to others, and make a contribution to the knowledge society beyond their immediate circle and life-spans
Funders and Policy Makers	Make evidence-based decisions Eliminate unnecessary duplication of work Get greater return on investment
Researchers	Have all data and tools easily available, increasing productivity Cross disciplinary boundaries, gaining new insights and producing new solutions 'Stand on the shoulders of giants'
Enterprise and Industry	Use the best available information for R&D, increasing productivity Create new knowledge, markets and job opportunities Provide a strong industrial and economic base for European prosperity Increase opportunities for mobility and knowledge exchange

28

RIDING THE WAVE How Europe can gain from the rising tide of scientific data

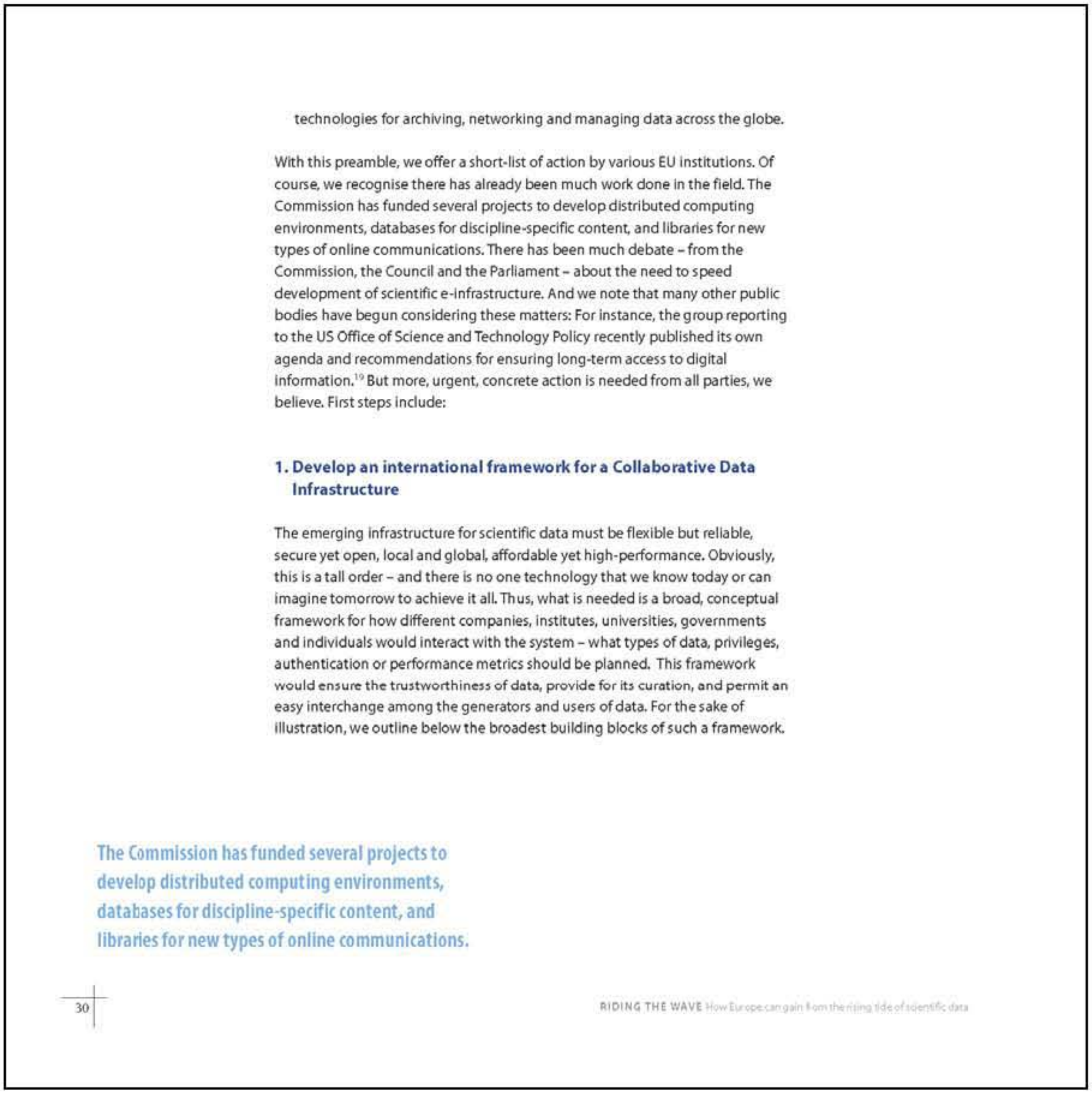


V. A call to action

The scientific and social benefits of our vision are numerous. But there are many other practical reasons to act. ICT is one of the main engines of economic growth. It is to our age what paved highways, national railroads and inter-continental telegraphs were to earlier generations. Yet in Europe, the industry underpinning this vital economic activity has had many difficulties. And, as the European Competiveness Council has noted, "the ICT impact on productivity growth is lower in the EU than in major trading partners."[18] A concerted European effort to build e-infrastructures for science will stimulate market demand for ICT. It will pull the best from ICT researchers, engineers and industrialists, spurring growth and jobs. And it will pave the digital highways that European science will need.

There is a clear role for government in all this. We urge our leaders to take into consideration the following:

- A good framework for the governance of data will be a source of strength in the most knowledge-intensive industries, fostering the growth of companies, goods, and services with the highest value-added. Those regions of the world that lead this policy debate, and develop the technologies and industry to support it, will gain competitive advantage.
- Scientific e-infrastructure is essential if we are to address the Grand Challenges of today. Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all fiendishly complex, cross-disciplinary problems that require high-performance data storage, smart analytics, transmission and mining to solve.
- Social cohesion will depend in part on how fairly and openly knowledge and information flow within our region and between the public and private sector. If information is power in the knowledge economy, governments must ensure that the benefits are appropriately distributed. Governments must work effectively through public-private partnerships to develop e-infrastructure.
- International collaboration is essential; there is no such thing as a purely local or national network anymore. We must collaborate in global architectures and governance for e-infrastructure, and we must share costs and



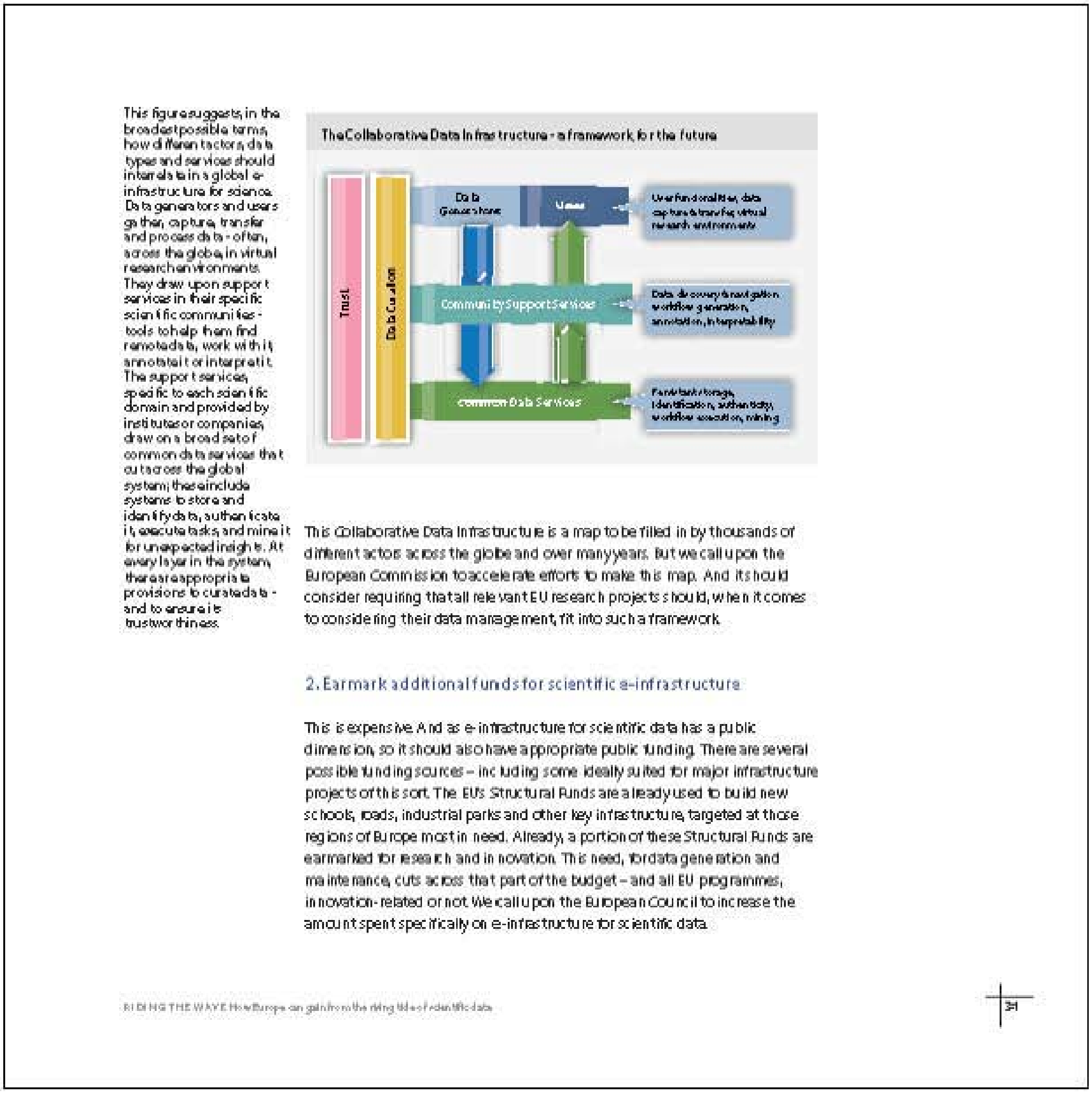
technologies for archiving, networking and managing data across the globe.

With this preamble, we offer a short-list of action by various EU institutions. Of course, we recognise there has already been much work done in the field. The Commission has funded several projects to develop distributed computing environments, databases for discipline-specific content, and libraries for new types of online communications. There has been much debate – from the Commission, the Council and the Parliament – about the need to speed development of scientific e-infrastructure. And we note that many other public bodies have begun considering these matters: For instance, the group reporting to the US Office of Science and Technology Policy recently published its own agenda and recommendations for ensuring long-term access to digital information.[19] But more, urgent, concrete action is needed from all parties, we believe. First steps include:

1. Develop an international framework for a Collaborative Data Infrastructure

The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance. Obviously, this is a tall order – and there is no one technology that we know today or can imagine tomorrow to achieve it all. Thus, what is needed is a broad, conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system – what types of data, privileges, authentication or performance metrics should be planned. This framework would ensure the trustworthiness of data, provide for its curation, and permit an easy interchange among the generators and users of data. For the sake of illustration, we outline below the broadest building blocks of such a framework.

The Commission has funded several projects to develop distributed computing environments, databases for discipline-specific content, and libraries for new types of online communications.



This figure suggests, in the broadest possible terms, how different actors, data types and services should interrelate in a global e-infrastructure for science. Data generators and users gather, capture, transfer and process data - often, across the globe, in virtual research environments. They draw upon support services in their specific scientific communities - tools to help them find remote data, work with it, annotate it or interpret it. The support services, specific to each scientific domain and provided by institutes or companies, draw on a broad set of common data services that cut across the global system; these include systems to store and identify data, authenticate it, execute tasks, and mine it for unexpected insights. At every layer in the system, there are appropriate provisions to curate data - and to ensure its trustworthiness.

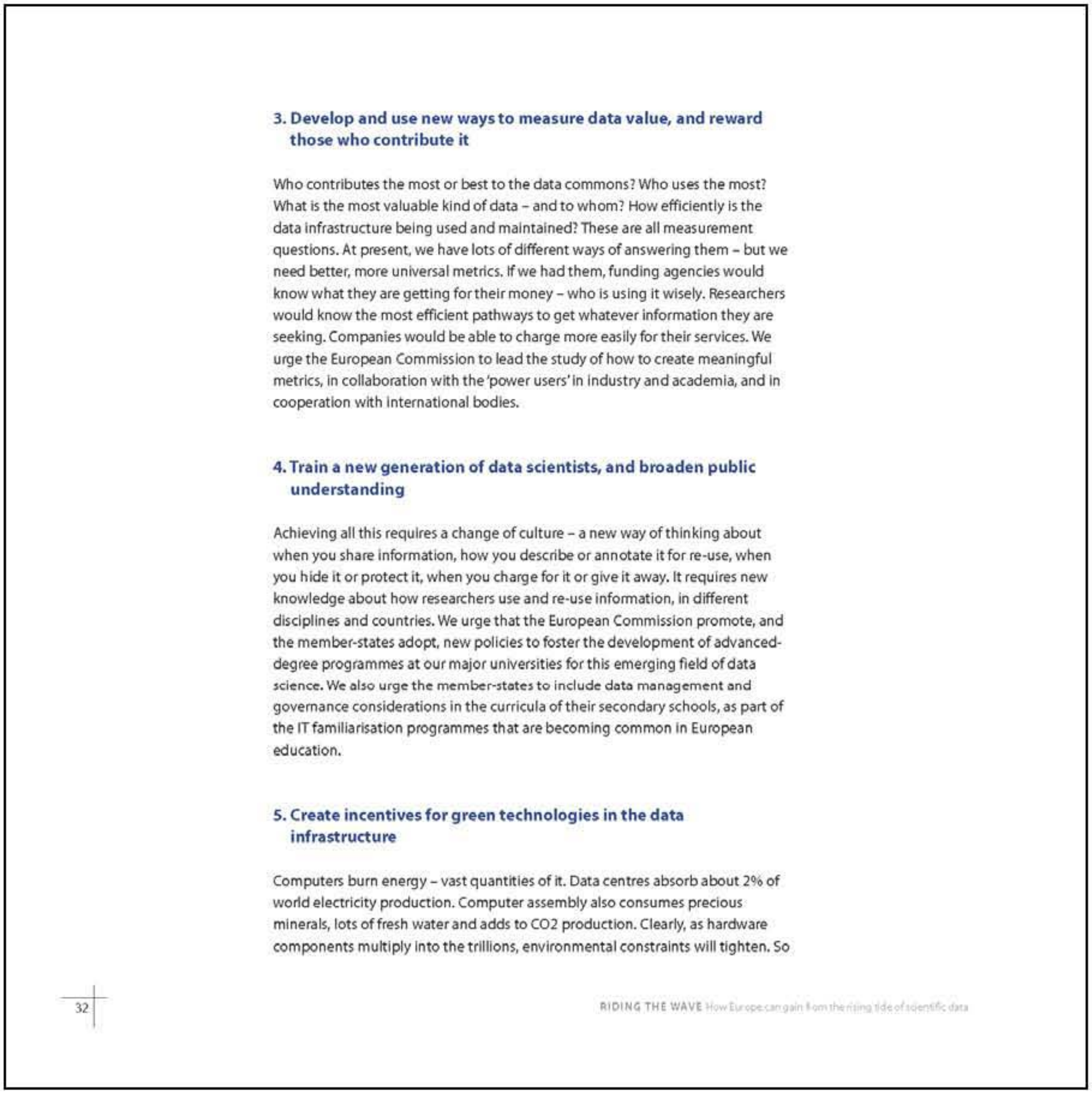
The Collaborative Data Infrastructure - a framework for the future

Trust
Data Curation
Data Generators
Users
User functionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation, workflow generation, annotation, interpretability
Common Data Services
Persistent storage, identification, authenticity, workflow execution, mining

This Collaborative Data Infrastructure is a map to be filled in by thousands of different actors across the globe and over many years. But we call upon the European Commission to accelerate efforts to make this map. And it should consider requiring that all relevant EU research projects should, when it comes to considering their data management, fit into such a framework.

2. Earmark additional funds for scientific e-infrastructure

This is expensive. And as e-infrastructure for scientific data has a public dimension, so it should also have appropriate public funding. There are several possible funding sources – including some ideally suited for major infrastructure projects of this sort. The EU's Structural Funds are already used to build new schools, roads, industrial parks and other key infrastructure, targeted at those regions of Europe most in need. Already, a portion of these Structural Funds are earmarked for research and innovation. This need, for data generation and maintenance, cuts across that part of the budget – and all EU programmes, innovation-related or not. We call upon the European Council to increase the amount spent specifically on e-infrastructure for scientific data.



3. Develop and use new ways to measure data value, and reward those who contribute it

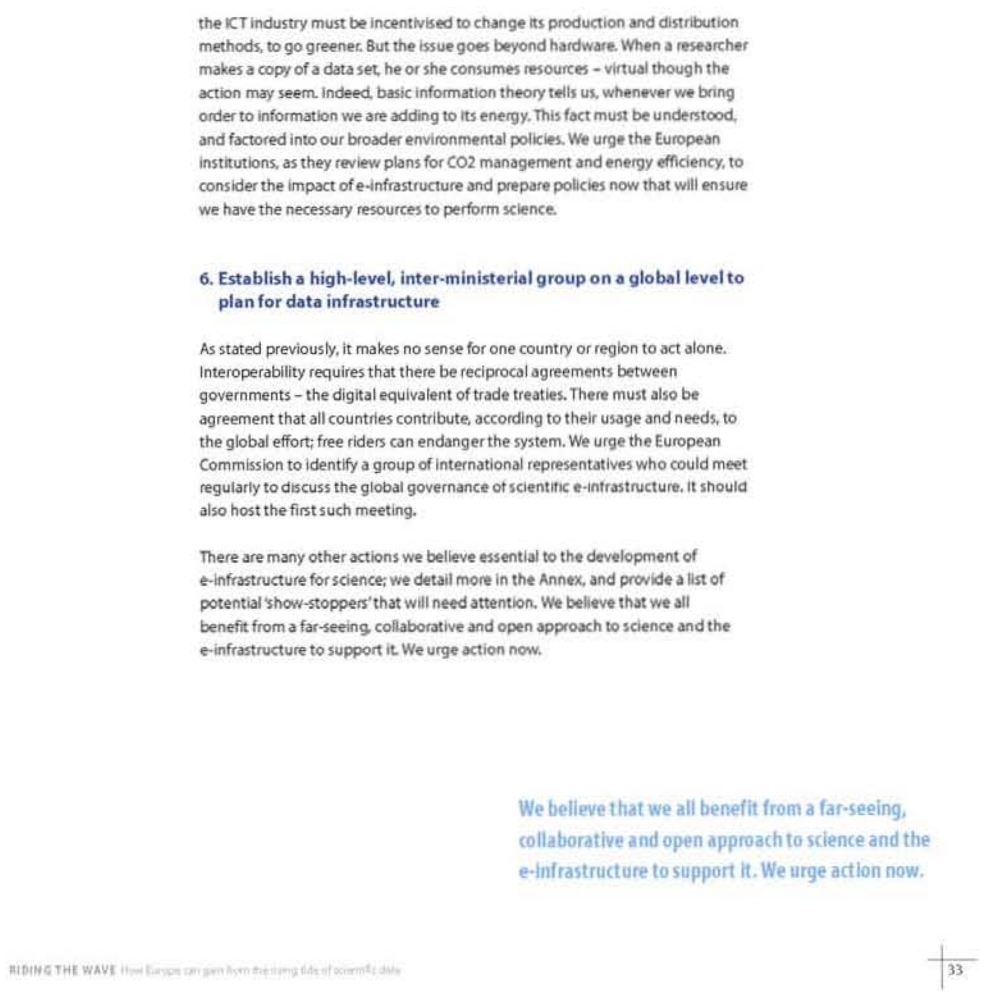
Who contributes the most or best to the data commons? Who uses the most? What is the most valuable kind of data – and to whom? How efficiently is the data infrastructure being used and maintained? These are all measurement questions. At present, we have lots of different ways of answering them – but we need better, more universal metrics. If we had them, funding agencies would know what they are getting for their money – who is using it wisely. Researchers would know the most efficient pathways to get whatever information they are seeking. Companies would be able to charge more easily for their services. We urge the European Commission to lead the study of how to create meaningful metrics, in collaboration with the 'power users' in industry and academia, and in cooperation with international bodies.

4. Train a new generation of data scientists, and broaden public understanding

Achieving all this requires a change of culture – a new way of thinking about when you share information, how you describe or annotate it for re-use, when you hide it or protect it, when you charge for it or give it away. It requires new knowledge about how researchers use and re-use information, in different disciplines and countries. We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for this emerging field of data science. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education.

5. Create incentives for green technologies in the data infrastructure

Computers burn energy – vast quantities of it. Data centres absorb about 2% of world electricity production. Computer assembly also consumes precious minerals, lots of fresh water and adds to CO2 production. Clearly, as hardware components multiply into the trillions, environmental constraints will tighten. So



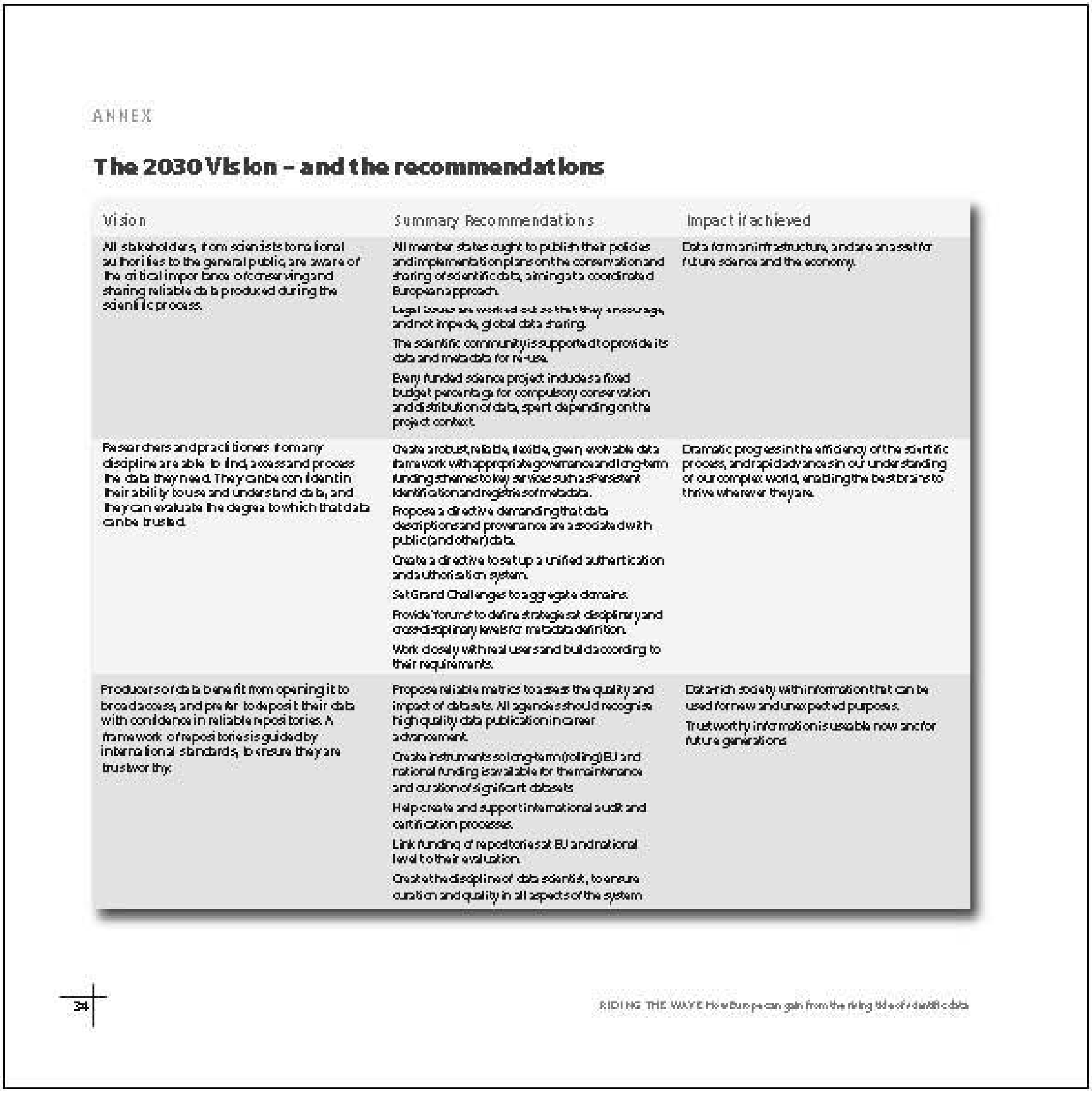
the ICT industry must be incentivised to change its production and distribution methods, to go greener. But the issue goes beyond hardware. When a researcher makes a copy of a data set, he or she consumes resources – virtual though the action may seem. Indeed, basic information theory tells us, whenever we bring order to information we are adding to its energy. This fact must be understood, and factored into our broader environmental policies. We urge the European institutions, as they review plans for CO2 management and energy efficiency, to consider the impact of e-infrastructure and prepare policies now that will ensure we have the necessary resources to perform science.

6. Establish a high-level, inter-ministerial group on a global level to plan for data infrastructure

As stated previously, it makes no sense for one country or region to act alone. Interoperability requires that there be reciprocal agreements between governments – the digital equivalent of trade treaties. There must also be agreement that all countries contribute, according to their usage and needs, to the global effort; free riders can endanger the system. We urge the European Commission to identify a group of international representatives who could meet regularly to discuss the global governance of scientific e-infrastructure. It should also host the first such meeting.

There are many other actions we believe essential to the development of e-infrastructure for science; we detail more in the Annex, and provide a list of potential 'show-stoppers' that will need attention. We believe that we all benefit from a far-seeing, collaborative and open approach to science and the e-infrastructure to support it. We urge action now.

We believe that we all benefit from a far-seeing, collaborative and open approach to science and the e-infrastructure to support it. We urge action now.



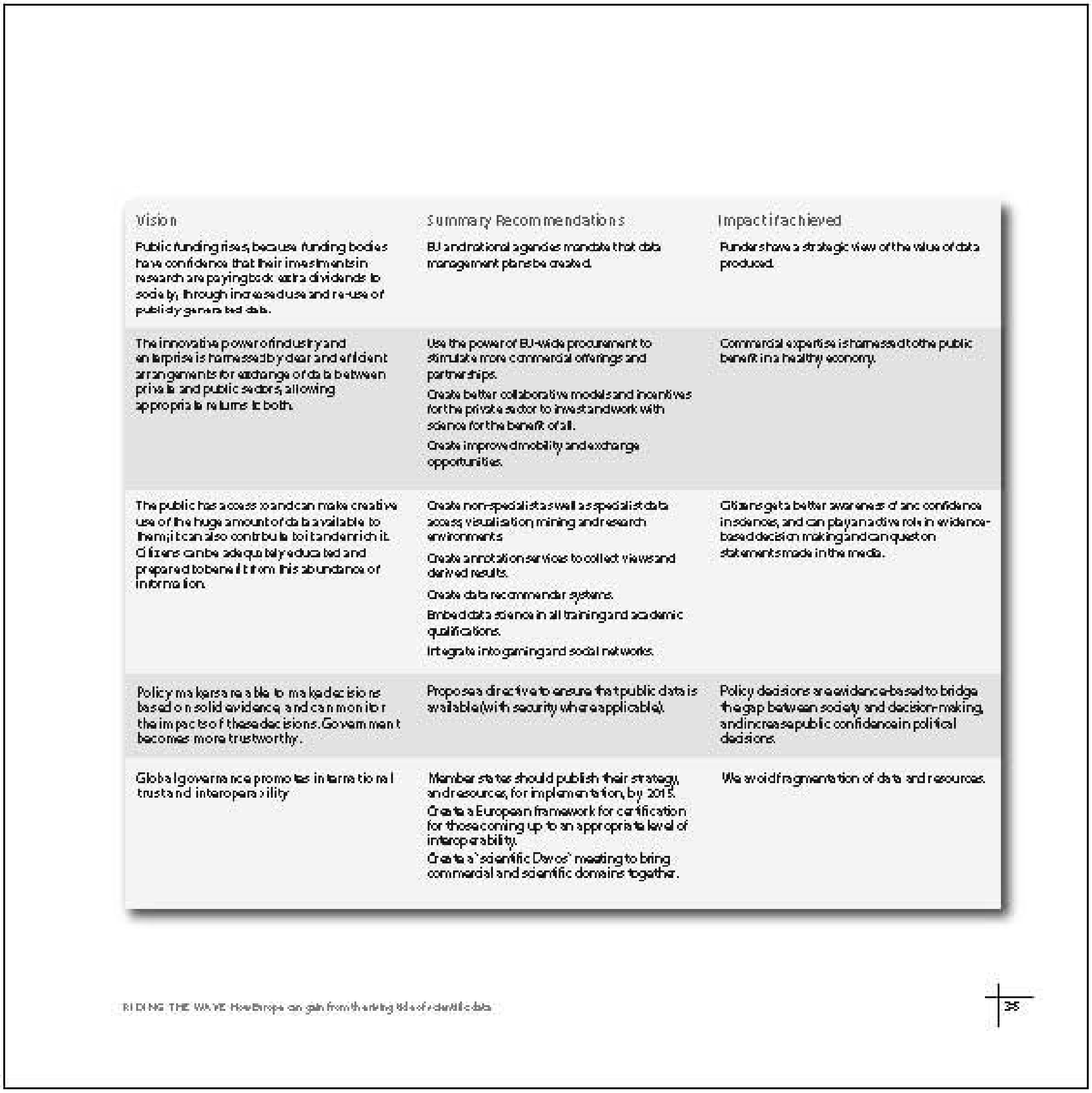
ANNEX

The 2030 Vision – and the recommendations

Vision	Summary Recommendations	Impact if achieved
All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.	All member states ought to publish their policies and implementation plans on the conservation and sharing of scientific data, aiming at a coordinated European approach. Legal issues are worked out so that they encourage, and not impede, global data sharing. The scientific community is supported to provide its data and metadata for re-use. Every funded science project includes a fixed budget percentage for compulsory conservation and distribution of data, spent depending on the project context.	Data form an infrastructure, and are an asset for future science and the economy.
Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.	Create a robust, reliable, flexible, green, evolvable data framework with appropriate governance and long-term funding schemes to key services such as Persistent Identification and registries of metadata. Propose a directive demanding that data descriptions and provenance are associated with public (and other) data. Create a directive to set up a unified authentication and authorisation system. Set Grand Challenges to aggregate domains. Provide 'forums' to define strategies at disciplinary and cross-disciplinary levels for metadata definition. Work closely with real users and build according to their requirements.	Dramatic progress in the efficiency of the scientific process, and rapid advances in our understanding of our complex world, enabling the best brains to thrive wherever they are.
Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy.	Propose reliable metrics to assess the quality and impact of datasets. All agencies should recognise high quality data publication in career advancement. Create instruments so long-term (rolling) EU and national funding is available for the maintenance and curation of significant datasets. Help create and support international audit and certification processes. Link funding of repositories at EU and national level to their evaluation. Create the discipline of data scientist, to ensure curation and quality in all aspects of the system.	Data-rich society with information that can be used for new and unexpected purposes. Trustworthy information is useable now and for future generations.

34

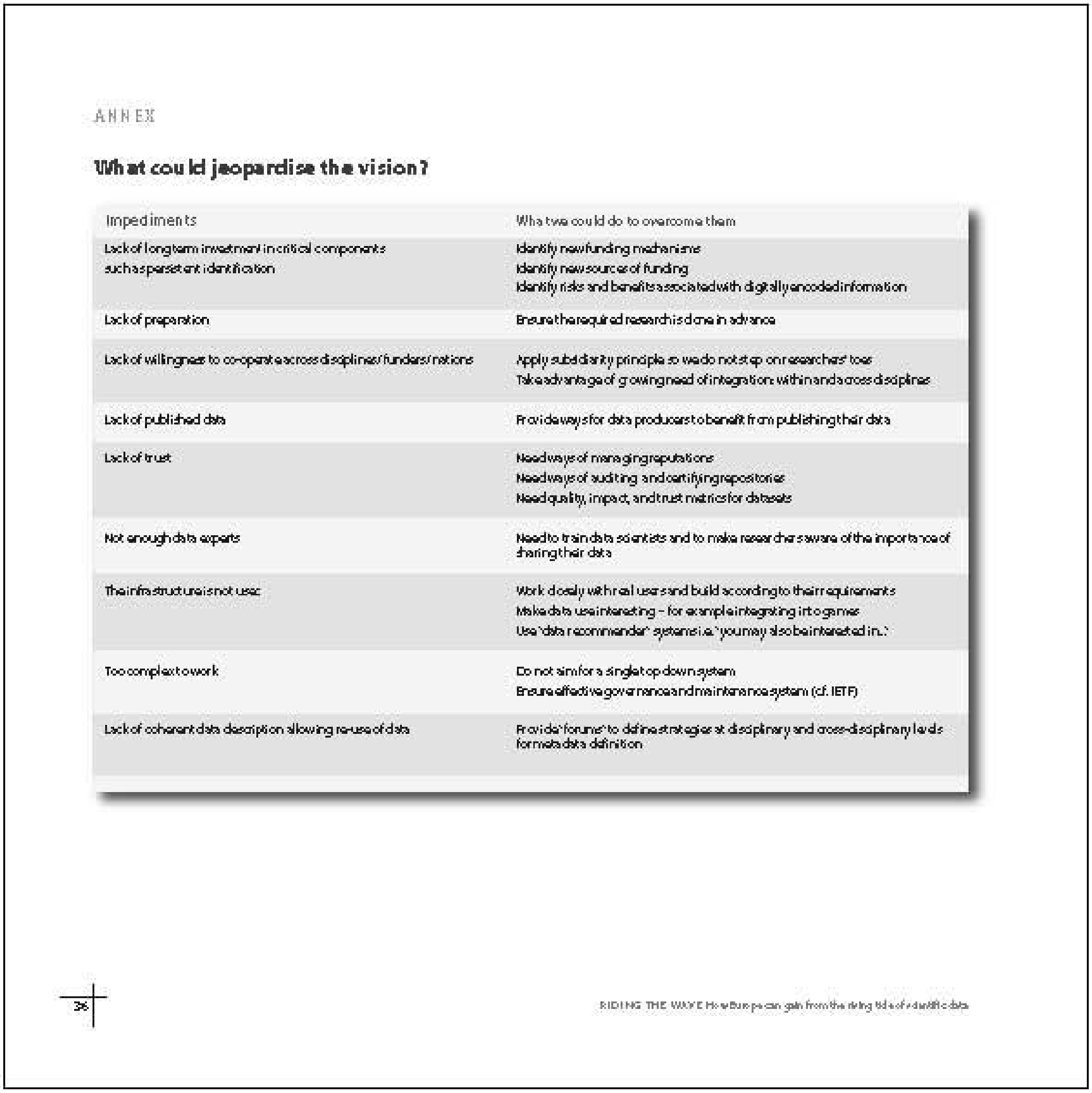
RIDING THE WAVE How Europe can gain from the rising tide of scientific data



Vision	Summary Recommendations	Impact if achieved
Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.	EU and national agencies mandate that data management plans be created.	Funders have a strategic view of the value of data produced.
The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.	Use the power of EU-wide procurement to stimulate more commercial offerings and partnerships. Create better collaborative models and incentives for the private sector to invest and work with science for the benefit of all. Create improved mobility and exchange opportunities.	Commercial expertise is harnessed to the public benefit in a healthy economy.
The public has access to and can make creative use of the huge amount of data available to them; it can also contribute to it and enrich it. Citizens can be adequately educated and prepared to benefit from this abundance of information.	Create non-specialist as well as specialist data access, visualisation, mining and research environments. Create annotation services to collect views and derived results. Create data recommender systems. Embed data science in all training and academic qualifications. Integrate into gaming and social networks.	Citizens get a better awareness of and confidence in sciences, and can play an active role in evidence-based decision making and can question statements made in the media.
Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.	Propose a directive to ensure that public data is available (with security where applicable).	Policy decisions are evidence-based to bridge the gap between society and decision-making, and increase public confidence in political decisions.
Global governance promotes international trust and interoperability	Member states should publish their strategy, and resources, for implementation, by 2015. Create a European framework for certification for those coming up to an appropriate level of interoperability. Create a 'scientific Davos' meeting to bring commercial and scientific domains together.	We avoid fragmentation of data and resources.

RIDING THE WAVE How Europe can gain from the rising tide of scientific data

35



ANNEX

What could jeopardise the vision?

Impediments	What we could do to overcome them
Lack of long term investment in critical components such as persistent identification	Identify new funding mechanisms Identify new sources of funding Identify risks and benefits associated with digitally encoded information
Lack of preparation	Ensure the required research is done in advance
Lack of willingness to co-operate across disciplines/funders/nations	Apply subsidiarity principle so we do not step on researchers' toes Take advantage of growing need of integration: within and across disciplines
Lack of published data	Provide ways for data producers to benefit from publishing their data
Lack of trust	Need ways of managing reputations Need ways of auditing and certifying repositories Need quality, impact, and trust metrics for datasets
Not enough data experts	Need to train data scientists and to make researchers aware of the importance of sharing their data
The infrastructure is not used	Work closely with real users and build according to their requirements Make data use interesting – for example integrating it into games Use "data recommender" systems i.e. "you may also be interested in..."
Too complex to work	Do not aim for a single top-down system Ensure effective governance and maintenance system (cf. IETF)
Lack of coherent data description allowing re-use of data	Provide "forums" to define strategies at disciplinary and cross-disciplinary levels for metadata definition

36

RIDING THE WAVE How Europe can gain from the rising tide of scientific data

About the High Level Group

The High Level Expert Group on Scientific Data was charged by the European Commission's Directorate-General for Information Society and Media to prepare a "vision 2030" for the evolution of e-infrastructure to scientific data.

Chair: **John Wood**, Secretary General of the Association of Commonwealth Universities

Thomas Andersson, Professor of Economics and former President, Jönköping University; Senior Advisor, Science, Technology and Innovation, Sultanate of Oman

Achim Bachem, Chairman, Board of Directors, Forschungszentrum Jülich GmbH

Christoph Best, European Bioinformatics Institute, Cambridge (UK)/Google UK Ltd, London (from September 2010).

Françoise Genova, Director, Strasbourg astronomical Data Centre; Observatoire Astronomique de Strasbourg, Université de Strasbourg/CNRS

Diego R. Lopez, RedIRIS

Wouter Los, Faculty of Science at the University of Amsterdam; Coordinator of preparatory project LifeWatch biodiversity research infrastructure; Vice Chair Governing Board of GBIF

Monica Marinucci, Director, Oracle Public Sector, Education and Research Business Unit

Laurent Romary, INRIA and Humboldt University

Herbert Van de Sompel, Staff Scientist, Los Alamos National Laboratory

Jens Vigen, Head Librarian, European Organization for Nuclear Research, CERN

Peter Wittenburg, Technical Director, Max Planck Institute for Psycholinguistics

Rapporteur: **David Giaretta**, STFC and Alliance for Permanent Access

Report Text: **Richard L. Hudson**, Science|Business

The HLG wishes to acknowledge the following individuals for their invaluable contribution to the discussions: Mirko Albani, Peter Doorn, Fabrizio Gagliardi, Daron Green, István Kenesei, Puneet Kishor, Kimmo Koski, Norbert Lossau, Linda Miller, Bernd Panzer-Steindel, Günter Stock, Ilkka Tuomi.

Design: Design4Science Ltd. Illustrations: Fletcher Ward Design

REFERENCES

[1] Hey, Tony; Stewart Tansley and Kristin Tolle, Eds. "*The Fourth Paradigm: Data-Intensive Scientific Discovery.*" Microsoft Research. Redmond, Wash. 2009. PDF at http://research.microsoft.com/enus/collaboration/fourthparadigm/

[2] Council of the European Union. "The future of ICT research, innovation and infrastructures - Adoption of Council Conclusions." 25 November 2009.

[3] Vinge, V. "The Creativity Machine". *Nature*, Vol. 440, March 2006.

[4] Hey, A.F.G. and A.E. Trefethen, in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G.C. Fox, A.J.G. Hey, Eds. Wiley, Hoboken, NJ, 2003.

[5,6] Beyea, Jan. "The Smart Electricity Grid and Scientific Research." *Science* 328: 979, 21 May 2010.

[7] "The Square Kilometre Array: Factsheet for Scientists and Engineers." SKA Program Development Office, April 2010. http://www.skatelescope.org/PDF/100420_SKA_Factsheet-Scientists-Engineers.pdf

[8] National Centre for Biotechnology Information. "What is GenBank?" http://www.ncbi.nlm.nih.gov/genbank/

[9] Institute for Systems Biology. "Systems biology – the 21st century science." http://www.systemsbiology.org

[10] The 1000 Genomes Project. http://www.1000genomes.org/

[11] Lofgren, Eric T. and Nina H. Fefferman. "The untapped potential of virtual game worlds to shed light on real world epidemics." *The Lancet Infectious Diseases*, VII:9 (625 – 629), September 2007.

[12] http://www.galaxyzoo.org/

[13] Irwin, A. "Constructing the Scientific Citizen: Science & Democracy in the Biosciences." *Public Understanding of Science* vol.10, pp.1-18 (2001).

[14] http://www.artportalen.se

[15] http://earth.esa.int/gscb/ltdp

[16] http://www.geant.net

[17] Survey results from the PARSE.Insight project (http://www.parse-insight.eu/) show the lack of awareness of preservation and reluctance to share data.

[18] Council of the European Union. *Ibid.*

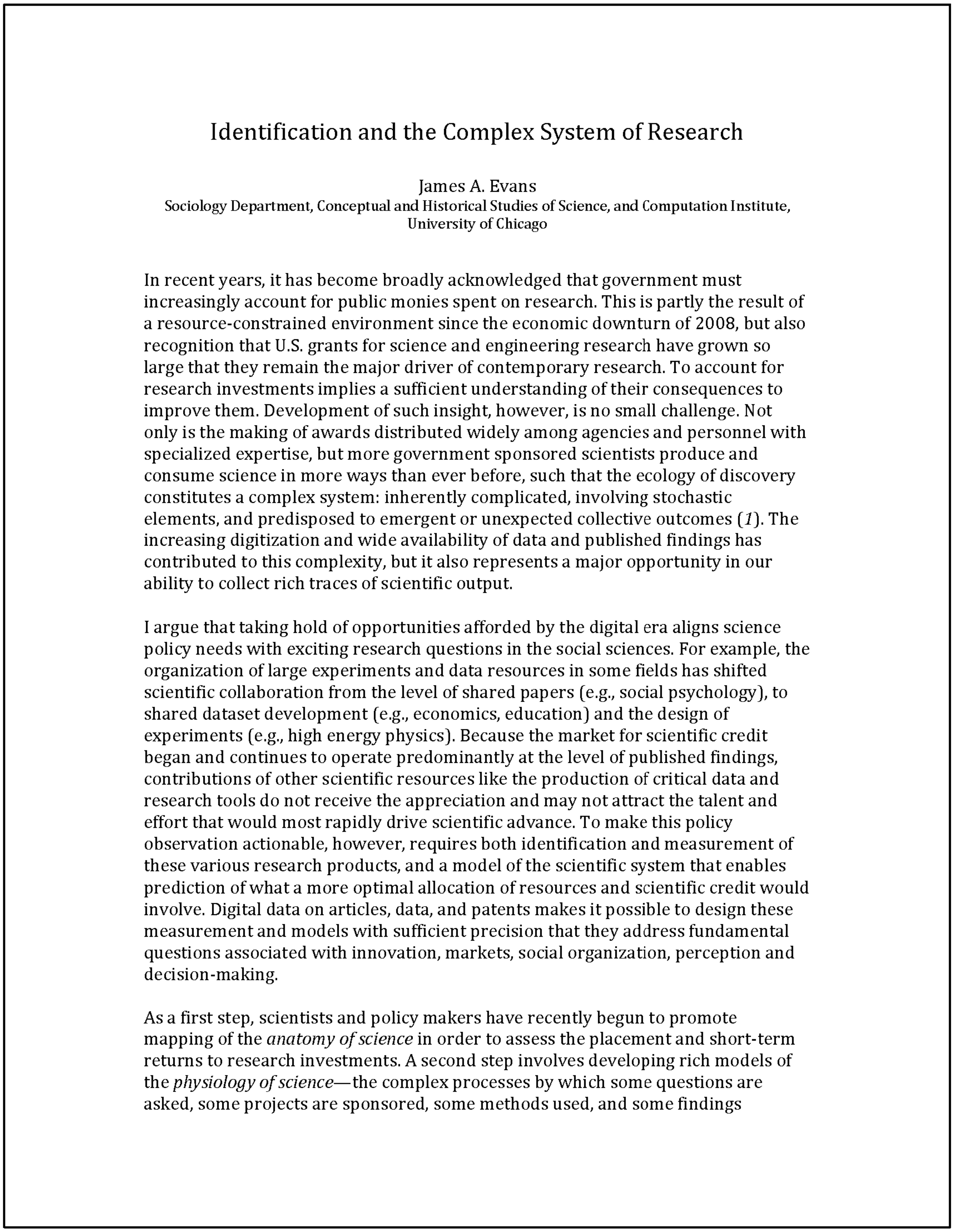
[19] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." February 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

For the 'Data Pyramid' graphic on page 18, the HLG wishes to acknowledge Berman, F. 2008. "Got data? A guide to data preservation in the information age." Communications of the ACM 51, 12 (Dec. 2008), 50-56. http://doi.acm.org/10.1145/1409360.1409376

The High Level Expert Group on Scientific Data was charged by the European Commission's Directorate-General for Information Society and Media to prepare a "vision 2030" for the evolution of e-infrastructure for scientific data. After meetings and consultations from December 2009 through June 2010, the group presents its outlook and recommendations.

e-infrastructure

Evans, James, “Identification and the Complex System of Research”



Identification and the Complex System of Research

James A. Evans
Sociology Department, Conceptual and Historical Studies of Science, and Computation Institute,
University of Chicago

In recent years, it has become broadly acknowledged that government must increasingly account for public monies spent on research. This is partly the result of a resource-constrained environment since the economic downturn of 2008, but also recognition that U.S. grants for science and engineering research have grown so large that they remain the major driver of contemporary research. To account for research investments implies a sufficient understanding of their consequences to improve them. Development of such insight, however, is no small challenge. Not only is the making of awards distributed widely among agencies and personnel with specialized expertise, but more government sponsored scientists produce and consume science in more ways than ever before, such that the ecology of discovery constitutes a complex system: inherently complicated, involving stochastic elements, and predisposed to emergent or unexpected collective outcomes (*1*). The increasing digitization and wide availability of data and published findings has contributed to this complexity, but it also represents a major opportunity in our ability to collect rich traces of scientific output.

I argue that taking hold of opportunities afforded by the digital era aligns science policy needs with exciting research questions in the social sciences. For example, the organization of large experiments and data resources in some fields has shifted scientific collaboration from the level of shared papers (e.g., social psychology), to shared dataset development (e.g., economics, education) and the design of experiments (e.g., high energy physics). Because the market for scientific credit began and continues to operate predominantly at the level of published findings, contributions of other scientific resources like the production of critical data and research tools do not receive the appreciation and may not attract the talent and effort that would most rapidly drive scientific advance. To make this policy observation actionable, however, requires both identification and measurement of these various research products, and a model of the scientific system that enables prediction of what a more optimal allocation of resources and scientific credit would involve. Digital data on articles, data, and patents makes it possible to design these measurement and models with sufficient precision that they address fundamental questions associated with innovation, markets, social organization, perception and decision-making.

As a first step, scientists and policy makers have recently begun to promote mapping of the *anatomy of science* in order to assess the placement and short-term returns to research investments. A second step involves developing rich models of the *physiology of science*—the complex processes by which some questions are asked, some projects are sponsored, some methods used, and some findings published, amplified and used in advance while others are not. To effectively address the first project hinges on the *identification* of essential elements in the research system, and the second on realistic *models* that capture essential interactions between those elements.

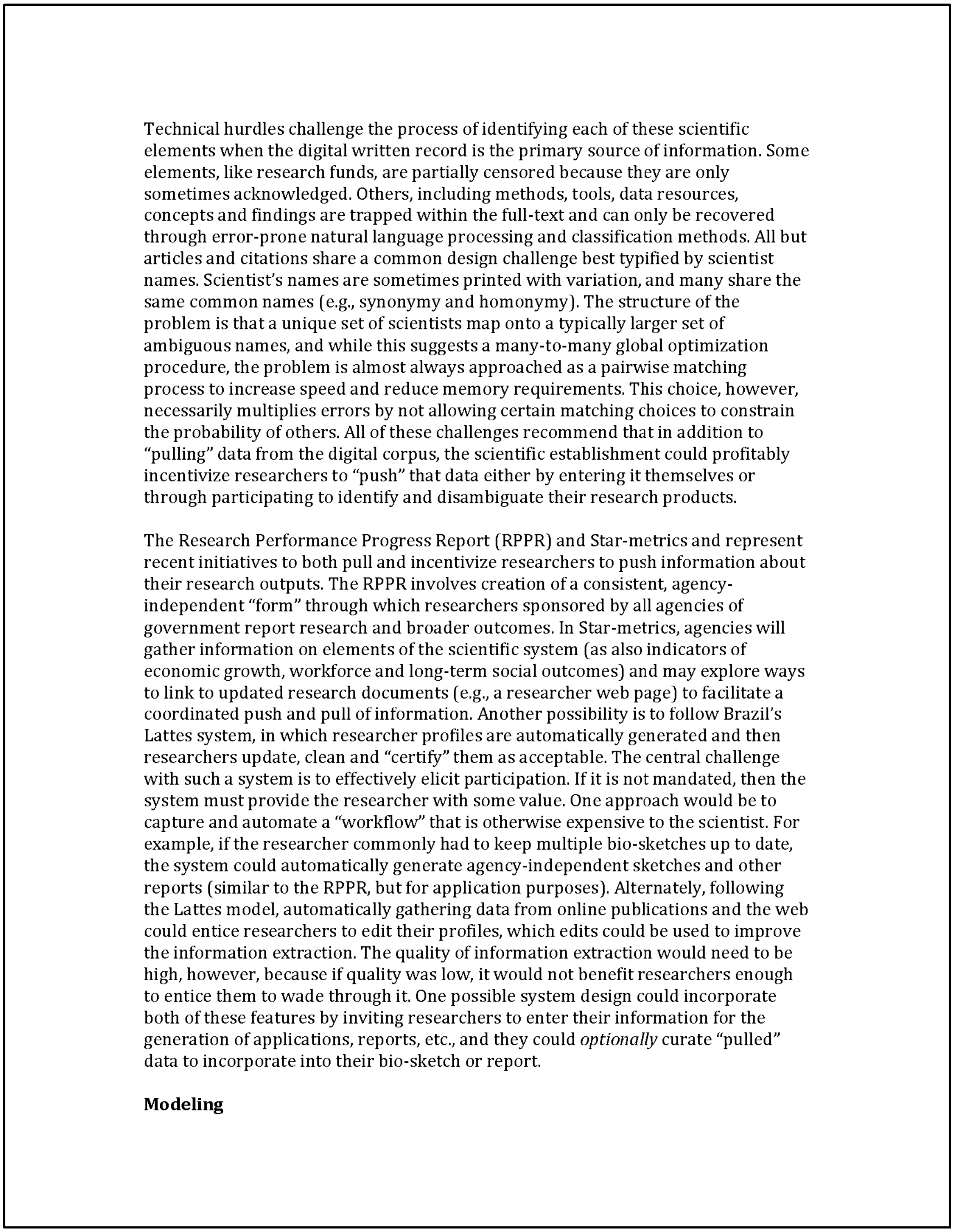
Identification and Measurement

The first step toward understanding the scientific system is to identify key elements in the system. These include, but are not restricted to the following:

1. Researchers (i.e., authors / inventors)
2. Research funds
3. Scientific knowledge:
    a. articles
    b. citations
    c. methods
    d. tools
    e. data resources
    f. concepts
    g. findings
4. Broader societal outcomes:
    a. Economic growth
        i. jobs
        ii. start-ups
        iii. patents
    b. Workforce
        i. student mobility into other jobs
        ii. student presence in jobs
    c. Long-term social outcomes
        i. health impact
        ii. environmental impact

The first two constitute research inputs, and third proximate outputs, which are themselves inputs to later stages of the research process. Although the first outputs—research documents and citations—have the most conceptual integrity and are the most often measured[1], they are unsatisfying as sole measurements because they do not represent the primary level of granularity at which scientists make "moves" and receive credit in science. Yes, academics publish and receive accolades for articles, but that is an outgrowth of their development, dissemination and promotion of methods, tools, data, concepts, and findings that seek to influence later work—to influence and advance science.

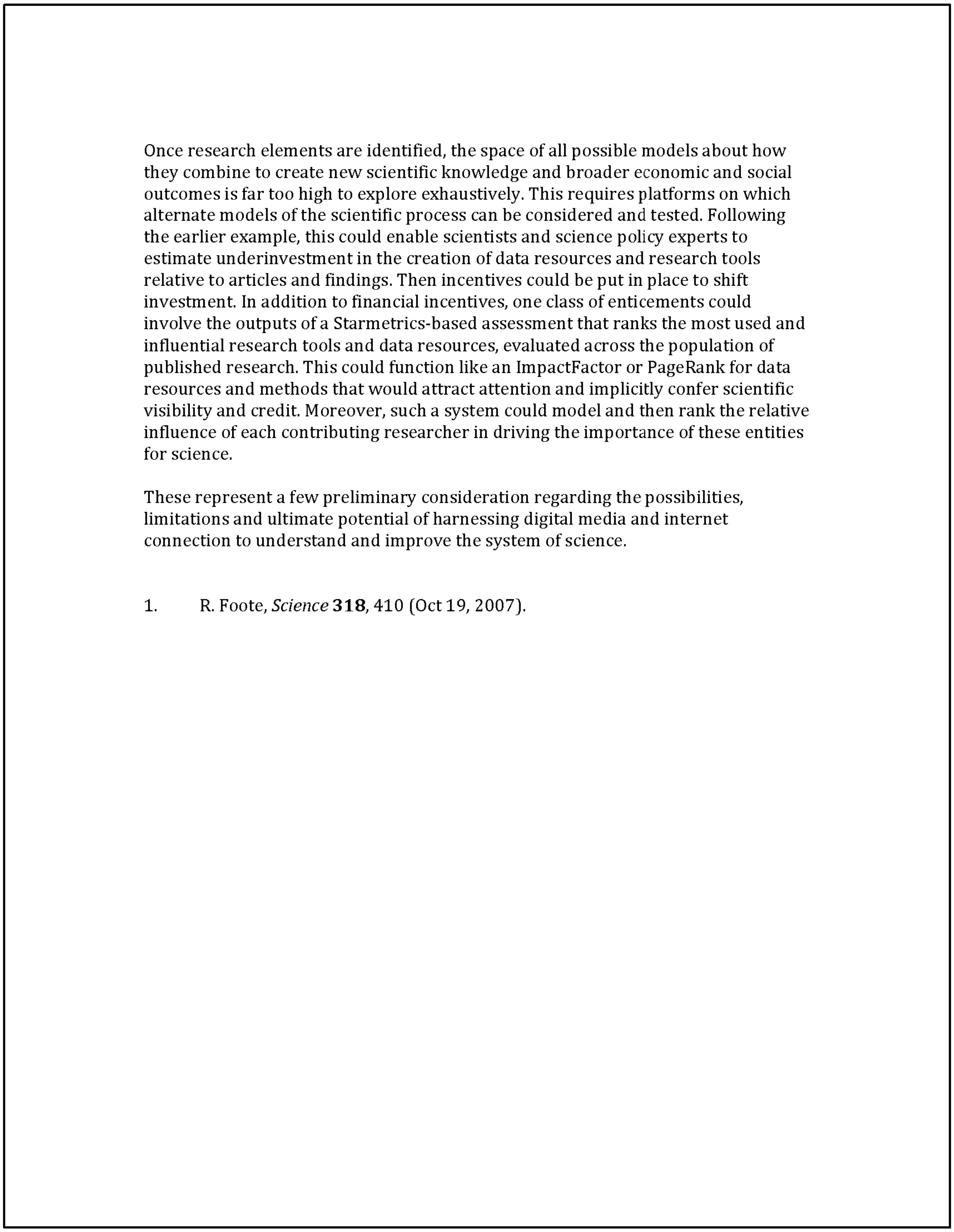
[1] Even the "integrity" of the article is beginning to change in the digital era with updating online books and papers.



Technical hurdles challenge the process of identifying each of these scientific elements when the digital written record is the primary source of information. Some elements, like research funds, are partially censored because they are only sometimes acknowledged. Others, including methods, tools, data resources, concepts and findings are trapped within the full-text and can only be recovered through error-prone natural language processing and classification methods. All but articles and citations share a common design challenge best typified by scientist names. Scientist's names are sometimes printed with variation, and many share the same common names (e.g., synonymy and homonymy). The structure of the problem is that a unique set of scientists map onto a typically larger set of ambiguous names, and while this suggests a many-to-many global optimization procedure, the problem is almost always approached as a pairwise matching process to increase speed and reduce memory requirements. This choice, however, necessarily multiplies errors by not allowing certain matching choices to constrain the probability of others. All of these challenges recommend that in addition to "pulling" data from the digital corpus, the scientific establishment could profitably incentivize researchers to "push" that data either by entering it themselves or through participating to identify and disambiguate their research products.

The Research Performance Progress Report (RPPR) and Star-metrics and represent recent initiatives to both pull and incentivize researchers to push information about their research outputs. The RPPR involves creation of a consistent, agency-independent "form" through which researchers sponsored by all agencies of government report research and broader outcomes. In Star-metrics, agencies will gather information on elements of the scientific system (as also indicators of economic growth, workforce and long-term social outcomes) and may explore ways to link to updated research documents (e.g., a researcher web page) to facilitate a coordinated push and pull of information. Another possibility is to follow Brazil's Lattes system, in which researcher profiles are automatically generated and then researchers update, clean and "certify" them as acceptable. The central challenge with such a system is to effectively elicit participation. If it is not mandated, then the system must provide the researcher with some value. One approach would be to capture and automate a "workflow" that is otherwise expensive to the scientist. For example, if the researcher commonly had to keep multiple bio-sketches up to date, the system could automatically generate agency-independent sketches and other reports (similar to the RPPR, but for application purposes). Alternately, following the Lattes model, automatically gathering data from online publications and the web could entice researchers to edit their profiles, which edits could be used to improve the information extraction. The quality of information extraction would need to be high, however, because if quality was low, it would not benefit researchers enough to entice them to wade through it. One possible system design could incorporate both of these features by inviting researchers to enter their information for the generation of applications, reports, etc., and they could *optionally* curate "pulled" data to incorporate into their bio-sketch or report.

Modeling

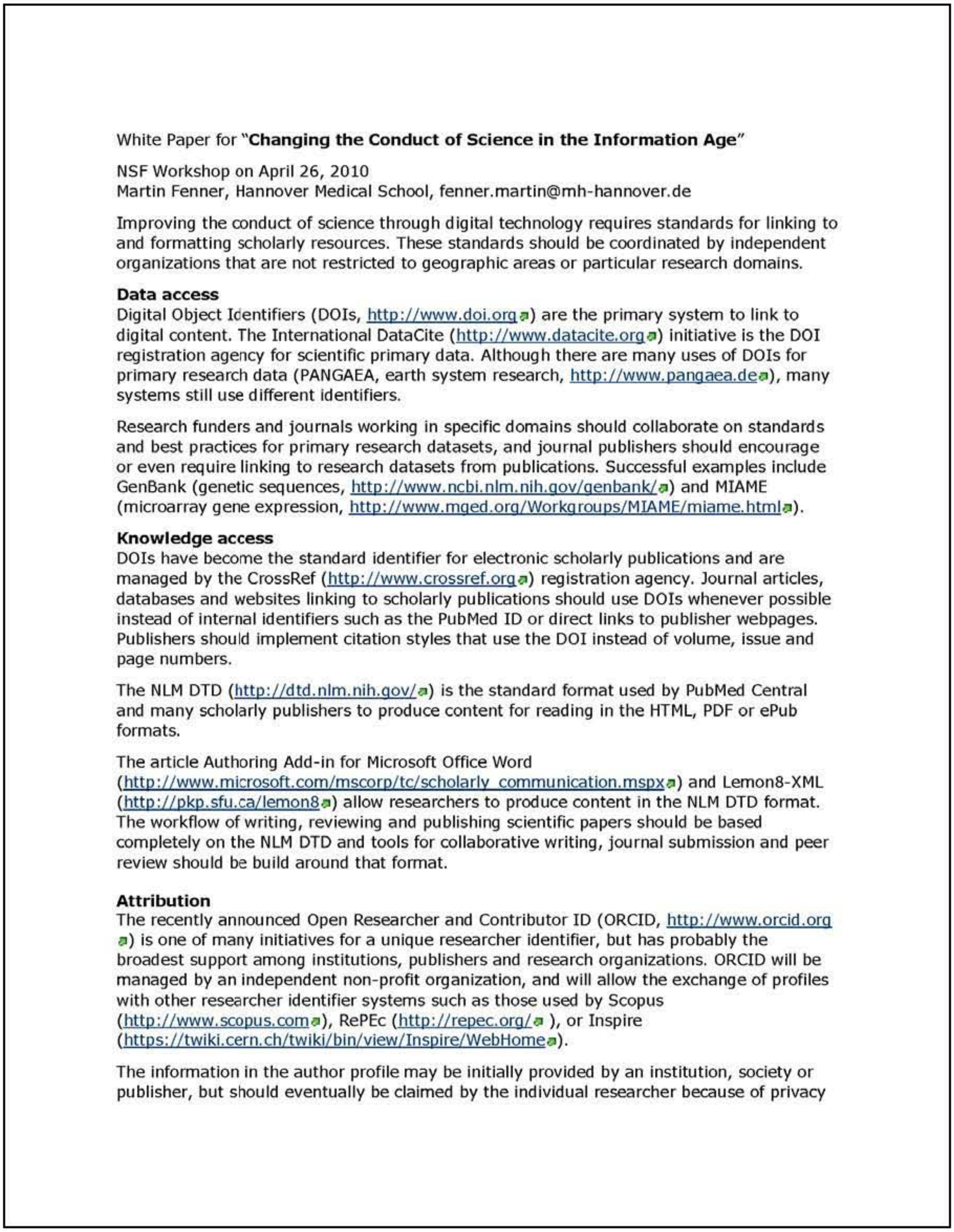


Once research elements are identified, the space of all possible models about how they combine to create new scientific knowledge and broader economic and social outcomes is far too high to explore exhaustively. This requires platforms on which alternate models of the scientific process can be considered and tested. Following the earlier example, this could enable scientists and science policy experts to estimate underinvestment in the creation of data resources and research tools relative to articles and findings. Then incentives could be put in place to shift investment. In addition to financial incentives, one class of enticements could involve the outputs of a Starmetrics-based assessment that ranks the most used and influential research tools and data resources, evaluated across the population of published research. This could function like an ImpactFactor or PageRank for data resources and methods that would attract attention and implicitly confer scientific visibility and credit. Moreover, such a system could model and then rank the relative influence of each contributing researcher in driving the importance of these entities for science.

These represent a few preliminary consideration regarding the possibilities, limitations and ultimate potential of harnessing digital media and internet connection to understand and improve the system of science.

1. R. Foote, *Science* **318**, 410 (Oct 19, 2007).

Fenner, Martin, "White Paper for Changing the Conduct of Science in the Information Age"



White Paper for "**Changing the Conduct of Science in the Information Age**"

NSF Workshop on April 26, 2010
Martin Fenner, Hannover Medical School, fenner.martin@mh-hannover.de

Improving the conduct of science through digital technology requires standards for linking to and formatting scholarly resources. These standards should be coordinated by independent organizations that are not restricted to geographic areas or particular research domains.

Data access

Digital Object Identifiers (DOIs, http://www.doi.org) are the primary system to link to digital content. The International DataCite (http://www.datacite.org) initiative is the DOI registration agency for scientific primary data. Although there are many uses of DOIs for primary research data (PANGAEA, earth system research, http://www.pangaea.de), many systems still use different identifiers.

Research funders and journals working in specific domains should collaborate on standards and best practices for primary research datasets, and journal publishers should encourage or even require linking to research datasets from publications. Successful examples include GenBank (genetic sequences, http://www.ncbi.nlm.nih.gov/genbank/) and MIAME (microarray gene expression, http://www.mged.org/Workgroups/MIAME/miame.html).

Knowledge access

DOIs have become the standard identifier for electronic scholarly publications and are managed by the CrossRef (http://www.crossref.org) registration agency. Journal articles, databases and websites linking to scholarly publications should use DOIs whenever possible instead of internal identifiers such as the PubMed ID or direct links to publisher webpages. Publishers should implement citation styles that use the DOI instead of volume, issue and page numbers.

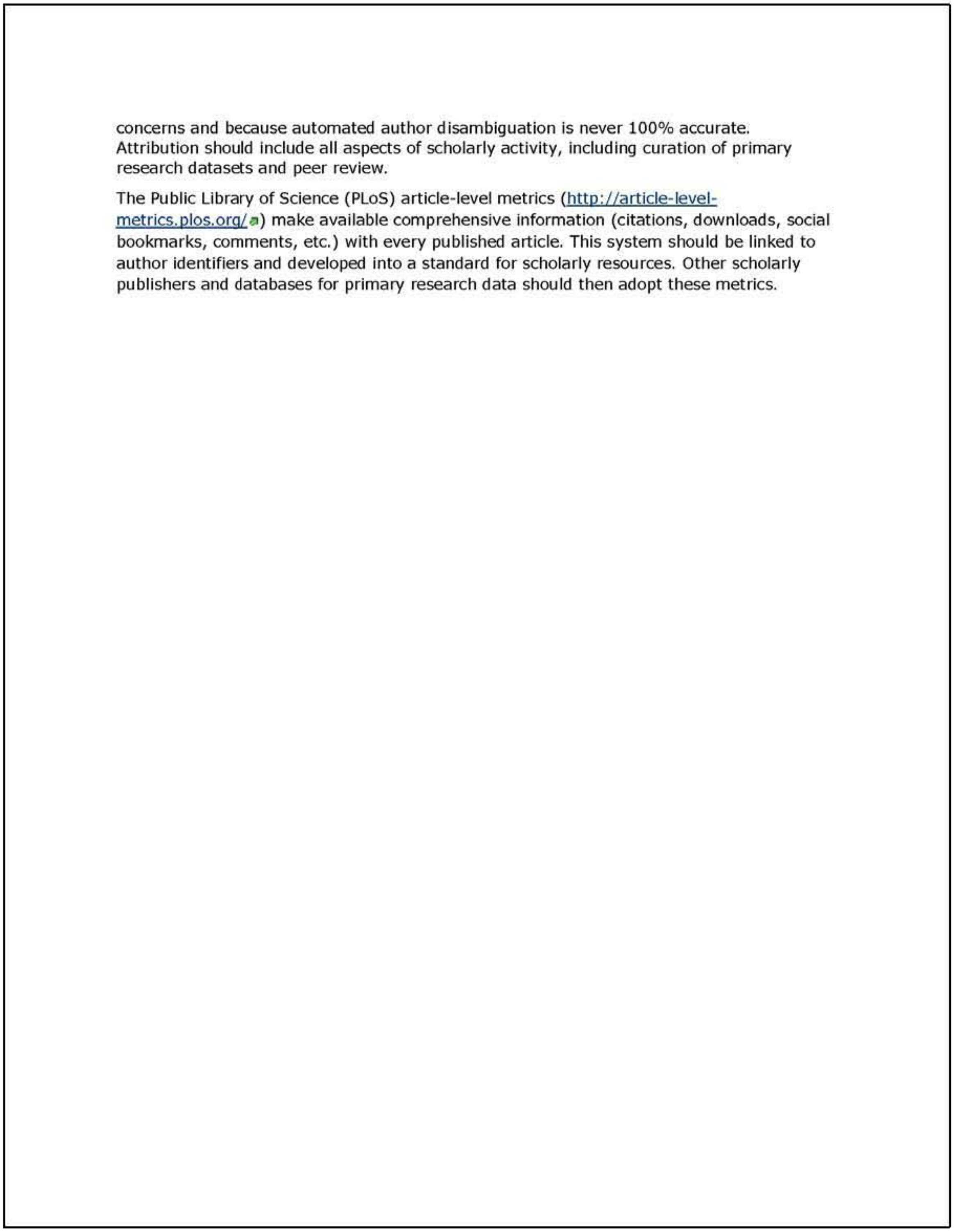
The NLM DTD (http://dtd.nlm.nih.gov/) is the standard format used by PubMed Central and many scholarly publishers to produce content for reading in the HTML, PDF or ePub formats.

The article Authoring Add-in for Microsoft Office Word (http://www.microsoft.com/mscorp/tc/scholarly_communication.mspx) and Lemon8-XML (http://pkp.sfu.ca/lemon8) allow researchers to produce content in the NLM DTD format. The workflow of writing, reviewing and publishing scientific papers should be based completely on the NLM DTD and tools for collaborative writing, journal submission and peer review should be build around that format.

Attribution

The recently announced Open Researcher and Contributor ID (ORCID, http://www.orcid.org) is one of many initiatives for a unique researcher identifier, but has probably the broadest support among institutions, publishers and research organizations. ORCID will be managed by an independent non-profit organization, and will allow the exchange of profiles with other researcher identifier systems such as those used by Scopus (http://www.scopus.com), RePEc (http://repec.org/), or Inspire (https://twiki.cern.ch/twiki/bin/view/Inspire/WebHome).

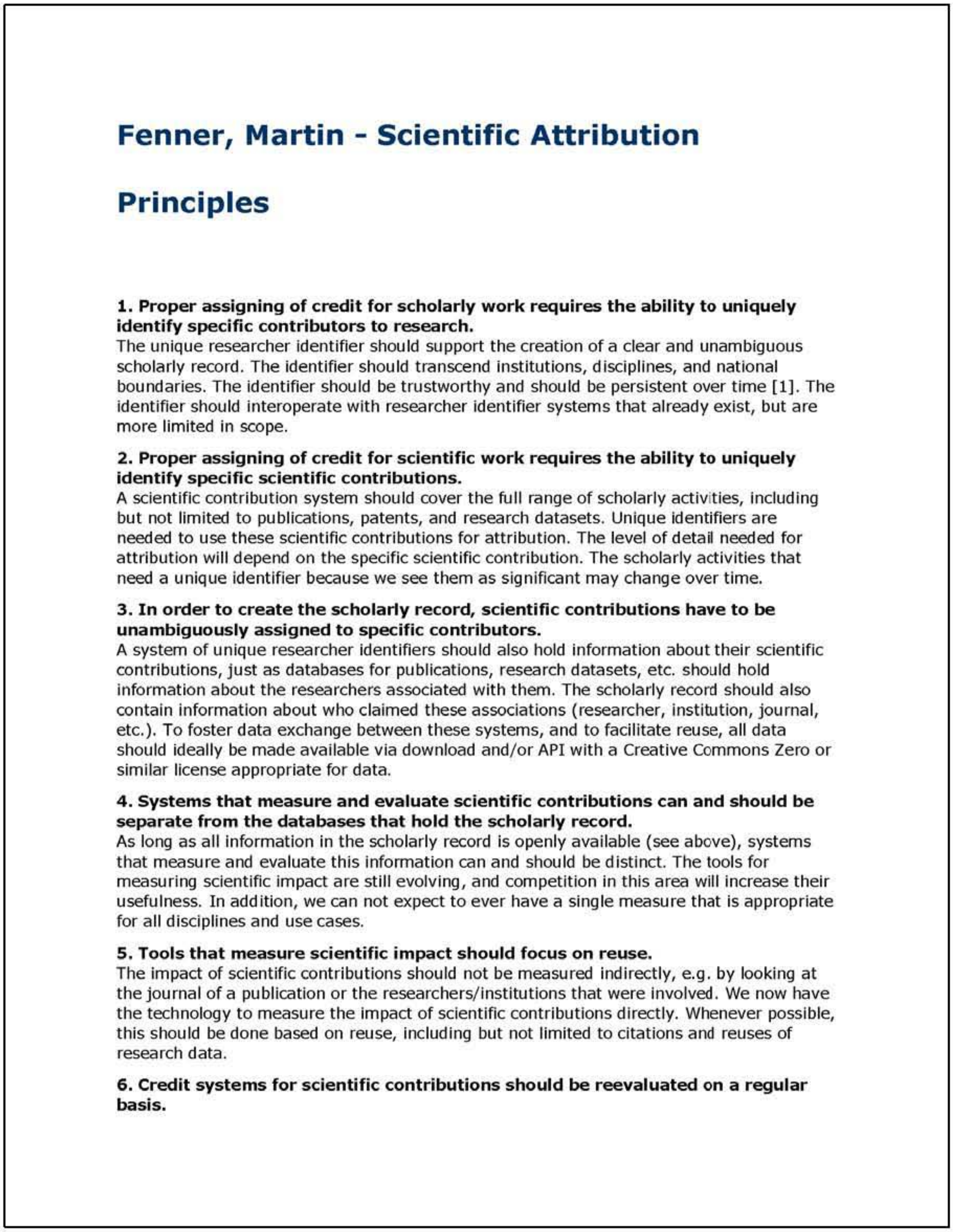
The information in the author profile may be initially provided by an institution, society or publisher, but should eventually be claimed by the individual researcher because of privacy



concerns and because automated author disambiguation is never 100% accurate. Attribution should include all aspects of scholarly activity, including curation of primary research datasets and peer review.

The Public Library of Science (PLoS) article-level metrics (http://article-level-metrics.plos.org/) make available comprehensive information (citations, downloads, social bookmarks, comments, etc.) with every published article. This system should be linked to author identifiers and developed into a standard for scholarly resources. Other scholarly publishers and databases for primary research data should then adopt these metrics.

Fenner, Martin, “Scientific Attribution Principles”



Fenner, Martin - Scientific Attribution Principles

1. Proper assigning of credit for scholarly work requires the ability to uniquely identify specific contributors to research.
The unique researcher identifier should support the creation of a clear and unambiguous scholarly record. The identifier should transcend institutions, disciplines, and national boundaries. The identifier should be trustworthy and should be persistent over time [1]. The identifier should interoperate with researcher identifier systems that already exist, but are more limited in scope.

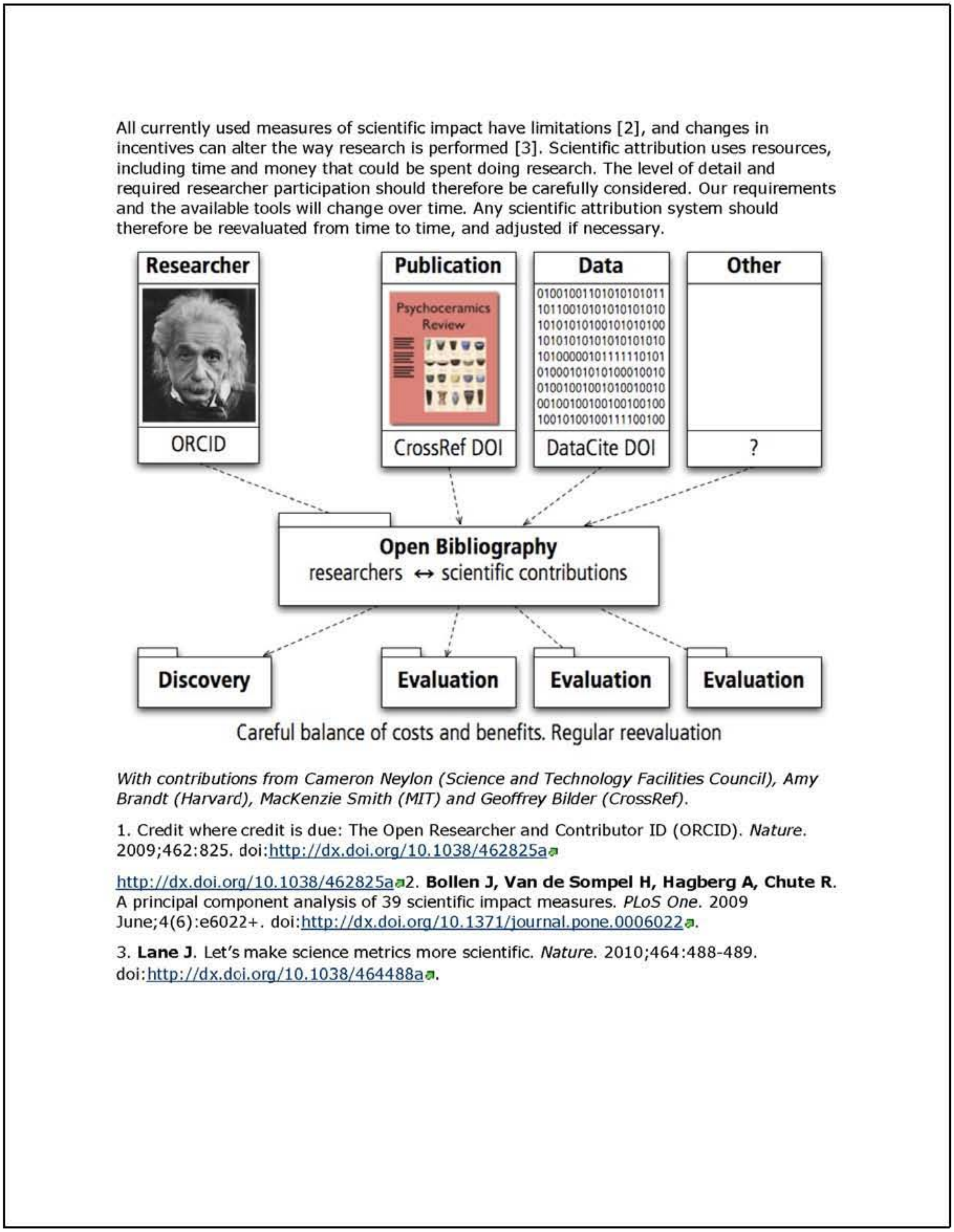
2. Proper assigning of credit for scientific work requires the ability to uniquely identify specific scientific contributions.
A scientific contribution system should cover the full range of scholarly activities, including but not limited to publications, patents, and research datasets. Unique identifiers are needed to use these scientific contributions for attribution. The level of detail needed for attribution will depend on the specific scientific contribution. The scholarly activities that need a unique identifier because we see them as significant may change over time.

3. In order to create the scholarly record, scientific contributions have to be unambiguously assigned to specific contributors.
A system of unique researcher identifiers should also hold information about their scientific contributions, just as databases for publications, research datasets, etc. should hold information about the researchers associated with them. The scholarly record should also contain information about who claimed these associations (researcher, institution, journal, etc.). To foster data exchange between these systems, and to facilitate reuse, all data should ideally be made available via download and/or API with a Creative Commons Zero or similar license appropriate for data.

4. Systems that measure and evaluate scientific contributions can and should be separate from the databases that hold the scholarly record.
As long as all information in the scholarly record is openly available (see above), systems that measure and evaluate this information can and should be distinct. The tools for measuring scientific impact are still evolving, and competition in this area will increase their usefulness. In addition, we can not expect to ever have a single measure that is appropriate for all disciplines and use cases.

5. Tools that measure scientific impact should focus on reuse.
The impact of scientific contributions should not be measured indirectly, e.g. by looking at the journal of a publication or the researchers/institutions that were involved. We now have the technology to measure the impact of scientific contributions directly. Whenever possible, this should be done based on reuse, including but not limited to citations and reuses of research data.

6. Credit systems for scientific contributions should be reevaluated on a regular basis.



All currently used measures of scientific impact have limitations [2], and changes in incentives can alter the way research is performed [3]. Scientific attribution uses resources, including time and money that could be spent doing research. The level of detail and required researcher participation should therefore be carefully considered. Our requirements and the available tools will change over time. Any scientific attribution system should therefore be reevaluated from time to time, and adjusted if necessary.

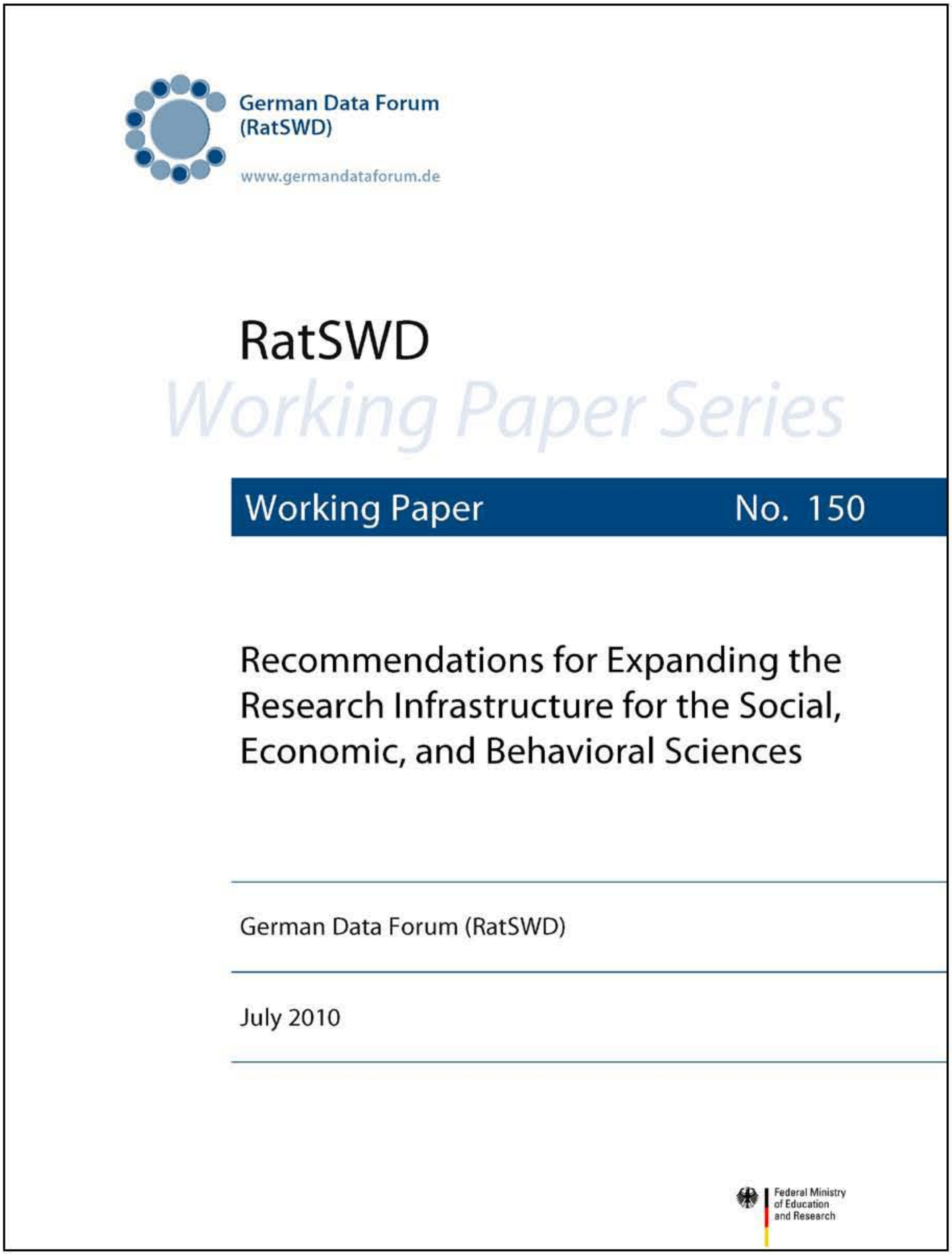
With contributions from Cameron Neylon (Science and Technology Facilities Council), Amy Brandt (Harvard), MacKenzie Smith (MIT) and Geoffrey Bilder (CrossRef).

1. Credit where credit is due: The Open Researcher and Contributor ID (ORCID). *Nature*. 2009;462:825. doi:http://dx.doi.org/10.1038/462825a

http://dx.doi.org/10.1038/462825a2. **Bollen J, Van de Sompel H, Hagberg A, Chute R**. A principal component analysis of 39 scientific impact measures. *PLoS One*. 2009 June;4(6):e6022+. doi:http://dx.doi.org/10.1371/journal.pone.0006022.

3. **Lane J**. Let's make science metrics more scientific. *Nature*. 2010;464:488-489. doi:http://dx.doi.org/10.1038/464488a.

German Data Forum, "RatSWD Working Paper Series No. 150: Recommendations for Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences"



German Data Forum
(RatSWD)

www.germandataforum.de

RatSWD
Working Paper Series

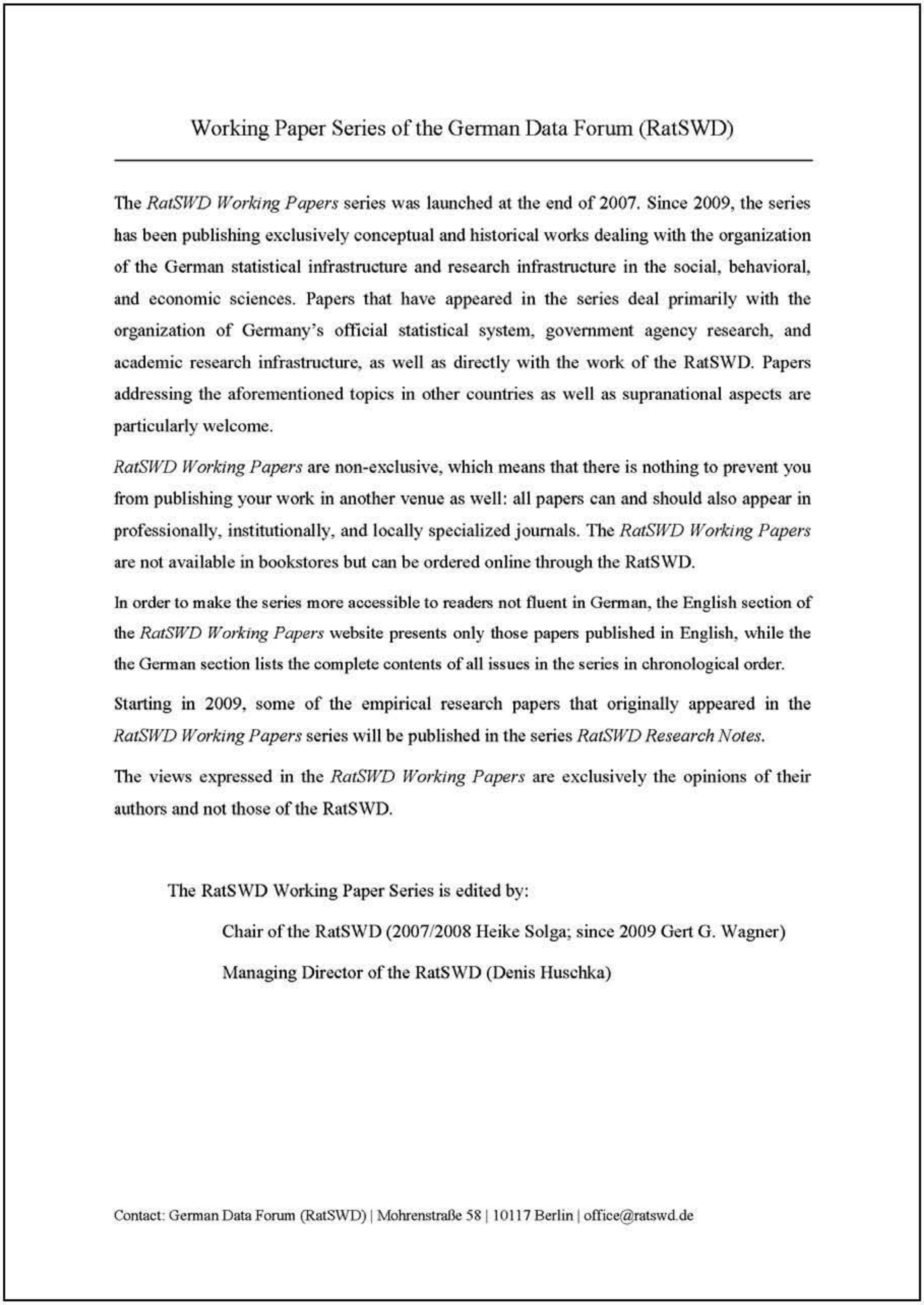
Working Paper No. 150

Recommendations for Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences

German Data Forum (RatSWD)

July 2010

Federal Ministry
of Education
and Research



Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

The RatSWD Working Paper Series is edited by:

Chair of the RatSWD (2007/2008 Heike Solga; since 2009 Gert G. Wagner)

Managing Director of the RatSWD (Denis Huschka)

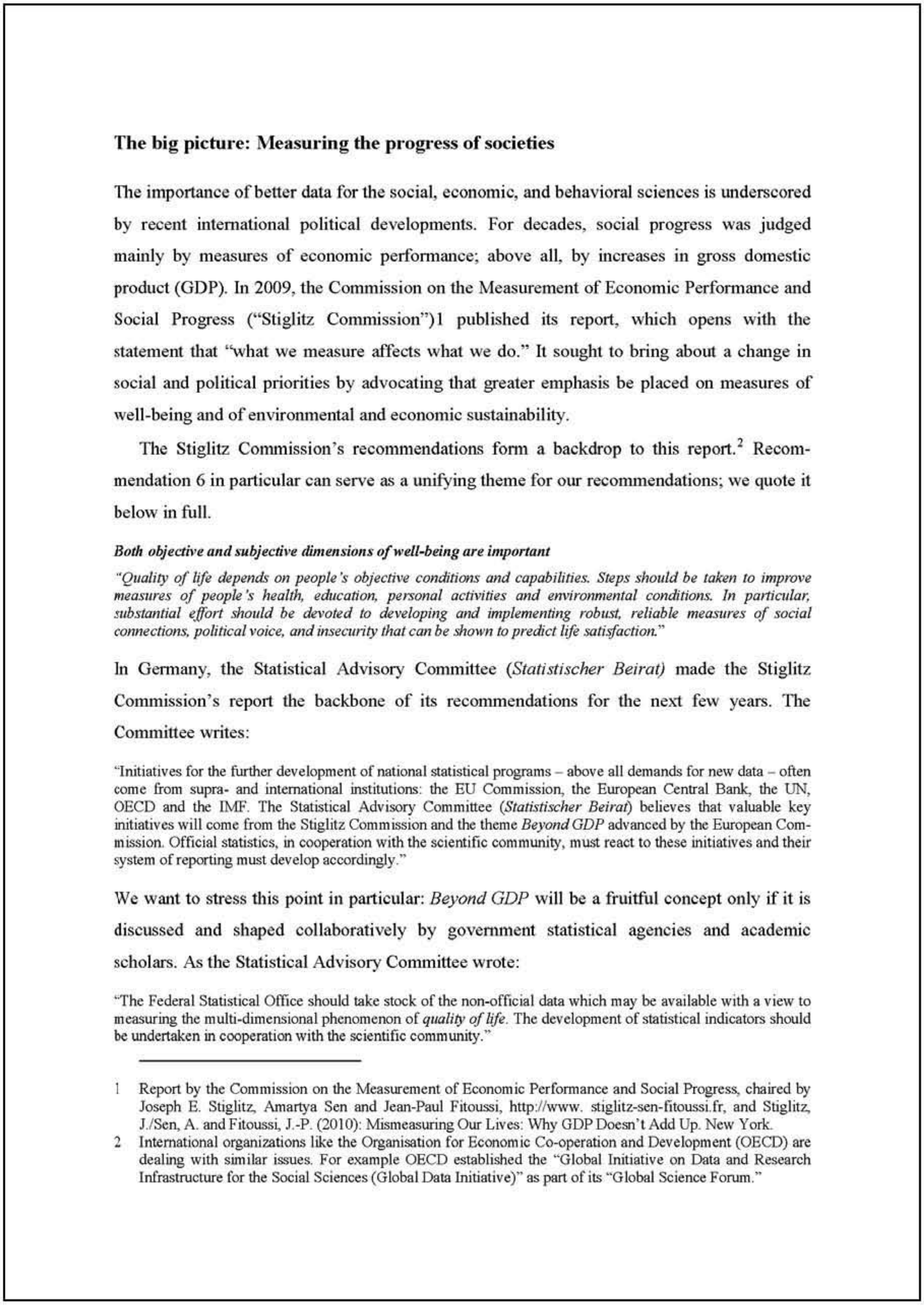
Contact: German Data Forum (RatSWD) | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

German Data Forum (RatSWD)[*]

Recommendations

For Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences

[*] The German Data Forum (RatSWD) adopted these recommendations at its 25th meeting on June 25, 2010, in Berlin. The recommendations will be published together with the underlying expert reports in a two-volume compendium: German Data Forum (RatSWD) (ed.), Building on Progress – Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Opladen, Budrich, 2010.



The big picture: Measuring the progress of societies

The importance of better data for the social, economic, and behavioral sciences is underscored by recent international political developments. For decades, social progress was judged mainly by measures of economic performance; above all, by increases in gross domestic product (GDP). In 2009, the Commission on the Measurement of Economic Performance and Social Progress ("Stiglitz Commission")1 published its report, which opens with the statement that "what we measure affects what we do." It sought to bring about a change in social and political priorities by advocating that greater emphasis be placed on measures of well-being and of environmental and economic sustainability.

The Stiglitz Commission's recommendations form a backdrop to this report.[2] Recommendation 6 in particular can serve as a unifying theme for our recommendations; we quote it below in full.

> ***Both objective and subjective dimensions of well-being are important***
>
> *"Quality of life depends on people's objective conditions and capabilities. Steps should be taken to improve measures of people's health, education, personal activities and environmental conditions. In particular, substantial effort should be devoted to developing and implementing robust, reliable measures of social connections, political voice, and insecurity that can be shown to predict life satisfaction."*

In Germany, the Statistical Advisory Committee (*Statistischer Beirat)* made the Stiglitz Commission's report the backbone of its recommendations for the next few years. The Committee writes:

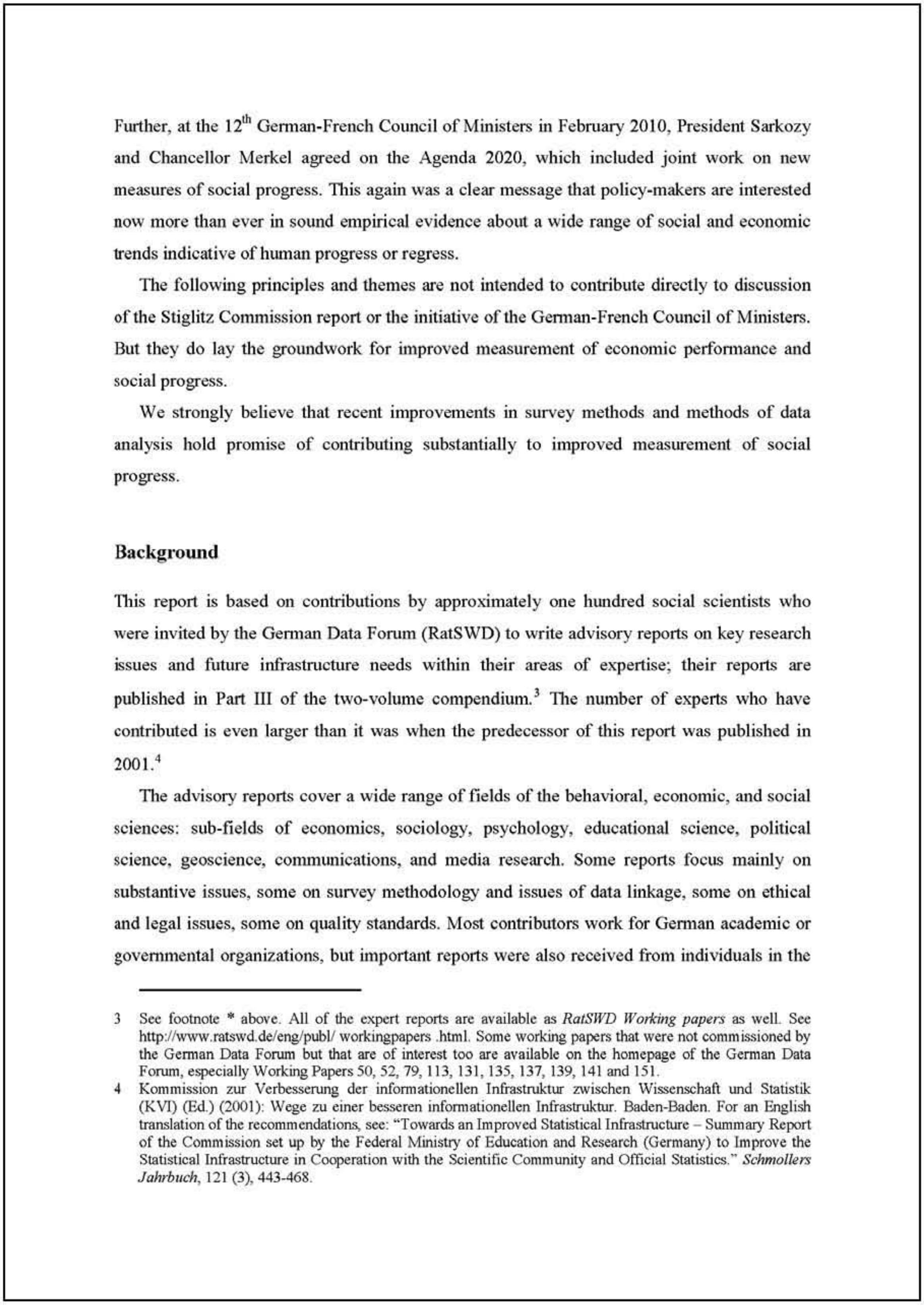
> "Initiatives for the further development of national statistical programs – above all demands for new data – often come from supra- and international institutions: the EU Commission, the European Central Bank, the UN, OECD and the IMF. The Statistical Advisory Committee (*Statistischer Beirat*) believes that valuable key initiatives will come from the Stiglitz Commission and the theme *Beyond GDP* advanced by the European Commission. Official statistics, in cooperation with the scientific community, must react to these initiatives and their system of reporting must develop accordingly."

We want to stress this point in particular: *Beyond GDP* will be a fruitful concept only if it is discussed and shaped collaboratively by government statistical agencies and academic scholars. As the Statistical Advisory Committee wrote:

> "The Federal Statistical Office should take stock of the non-official data which may be available with a view to measuring the multi-dimensional phenomenon of *quality of life*. The development of statistical indicators should be undertaken in cooperation with the scientific community."

[1] Report by the Commission on the Measurement of Economic Performance and Social Progress, chaired by Joseph E. Stiglitz, Amartya Sen and Jean-Paul Fitoussi, http://www. stiglitz-sen-fitoussi.fr, and Stiglitz, J./Sen, A. and Fitoussi, J.-P. (2010): Mismeasuring Our Lives: Why GDP Doesn't Add Up. New York.

[2] International organizations like the Organisation for Economic Co-operation and Development (OECD) are dealing with similar issues. For example OECD established the "Global Initiative on Data and Research Infrastructure for the Social Sciences (Global Data Initiative)" as part of its "Global Science Forum."



Further, at the 12th German-French Council of Ministers in February 2010, President Sarkozy and Chancellor Merkel agreed on the Agenda 2020, which included joint work on new measures of social progress. This again was a clear message that policy-makers are interested now more than ever in sound empirical evidence about a wide range of social and economic trends indicative of human progress or regress.

The following principles and themes are not intended to contribute directly to discussion of the Stiglitz Commission report or the initiative of the German-French Council of Ministers. But they do lay the groundwork for improved measurement of economic performance and social progress.

We strongly believe that recent improvements in survey methods and methods of data analysis hold promise of contributing substantially to improved measurement of social progress.

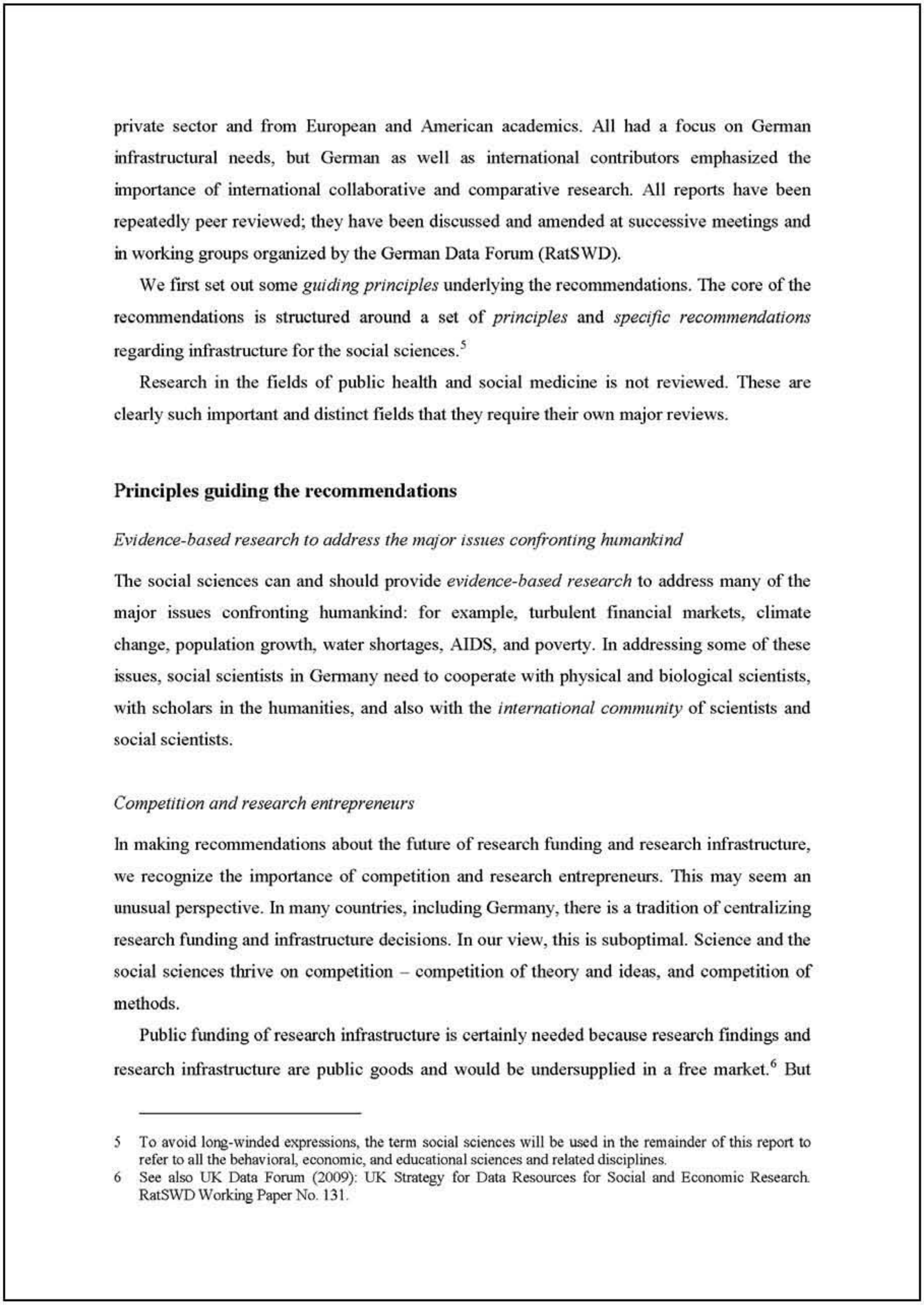
Background

This report is based on contributions by approximately one hundred social scientists who were invited by the German Data Forum (RatSWD) to write advisory reports on key research issues and future infrastructure needs within their areas of expertise; their reports are published in Part III of the two-volume compendium.[3] The number of experts who have contributed is even larger than it was when the predecessor of this report was published in 2001.[4]

The advisory reports cover a wide range of fields of the behavioral, economic, and social sciences: sub-fields of economics, sociology, psychology, educational science, political science, geoscience, communications, and media research. Some reports focus mainly on substantive issues, some on survey methodology and issues of data linkage, some on ethical and legal issues, some on quality standards. Most contributors work for German academic or governmental organizations, but important reports were also received from individuals in the

[3] See footnote * above. All of the expert reports are available as *RatSWD Working papers* as well. See http://www.ratswd.de/eng/publ/ workingpapers .html. Some working papers that were not commissioned by the German Data Forum but that are of interest too are available on the homepage of the German Data Forum, especially Working Papers 50, 52, 79, 113, 131, 135, 137, 139, 141 and 151.

[4] Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) (Ed.) (2001): Wege zu einer besseren informationellen Infrastruktur. Baden-Baden. For an English translation of the recommendations, see: "Towards an Improved Statistical Infrastructure – Summary Report of the Commission set up by the Federal Ministry of Education and Research (Germany) to Improve the Statistical Infrastructure in Cooperation with the Scientific Community and Official Statistics." *Schmollers Jahrbuch*, 121 (3), 443-468.



private sector and from European and American academics. All had a focus on German infrastructural needs, but German as well as international contributors emphasized the importance of international collaborative and comparative research. All reports have been repeatedly peer reviewed; they have been discussed and amended at successive meetings and in working groups organized by the German Data Forum (RatSWD).

We first set out some *guiding principles* underlying the recommendations. The core of the recommendations is structured around a set of *principles* and *specific recommendations* regarding infrastructure for the social sciences.[5]

Research in the fields of public health and social medicine is not reviewed. These are clearly such important and distinct fields that they require their own major reviews.

Principles guiding the recommendations

Evidence-based research to address the major issues confronting humankind

The social sciences can and should provide *evidence-based research* to address many of the major issues confronting humankind: for example, turbulent financial markets, climate change, population growth, water shortages, AIDS, and poverty. In addressing some of these issues, social scientists in Germany need to cooperate with physical and biological scientists, with scholars in the humanities, and also with the *international community* of scientists and social scientists.

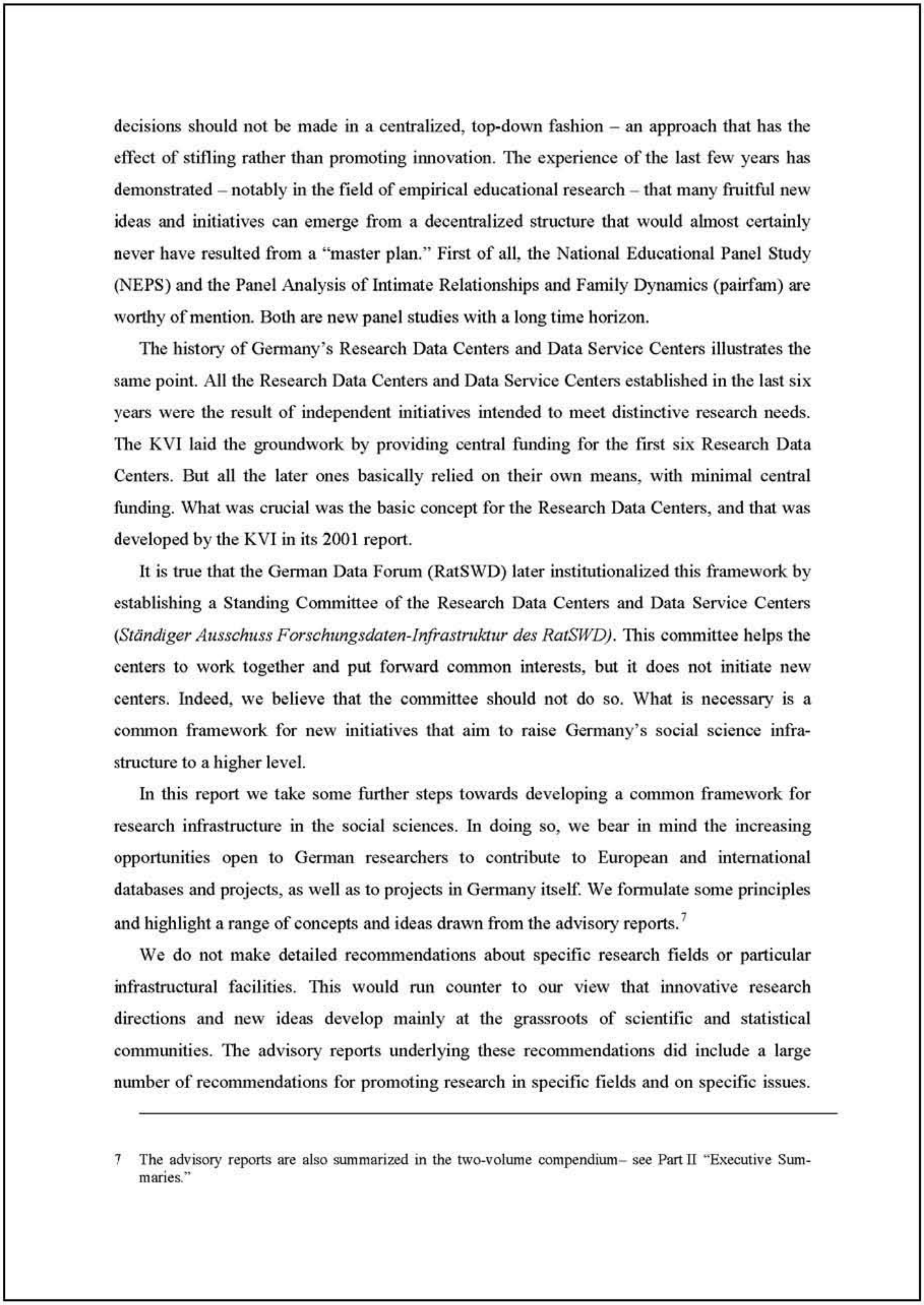
Competition and research entrepreneurs

In making recommendations about the future of research funding and research infrastructure, we recognize the importance of competition and research entrepreneurs. This may seem an unusual perspective. In many countries, including Germany, there is a tradition of centralizing research funding and infrastructure decisions. In our view, this is suboptimal. Science and the social sciences thrive on competition – competition of theory and ideas, and competition of methods.

Public funding of research infrastructure is certainly needed because research findings and research infrastructure are public goods and would be undersupplied in a free market.[6] But

[5] To avoid long-winded expressions, the term social sciences will be used in the remainder of this report to refer to all the behavioral, economic, and educational sciences and related disciplines.

[6] See also UK Data Forum (2009): UK Strategy for Data Resources for Social and Economic Research. RatSWD Working Paper No. 131.



decisions should not be made in a centralized, top-down fashion – an approach that has the effect of stifling rather than promoting innovation. The experience of the last few years has demonstrated – notably in the field of empirical educational research – that many fruitful new ideas and initiatives can emerge from a decentralized structure that would almost certainly never have resulted from a "master plan." First of all, the National Educational Panel Study (NEPS) and the Panel Analysis of Intimate Relationships and Family Dynamics (pairfam) are worthy of mention. Both are new panel studies with a long time horizon.

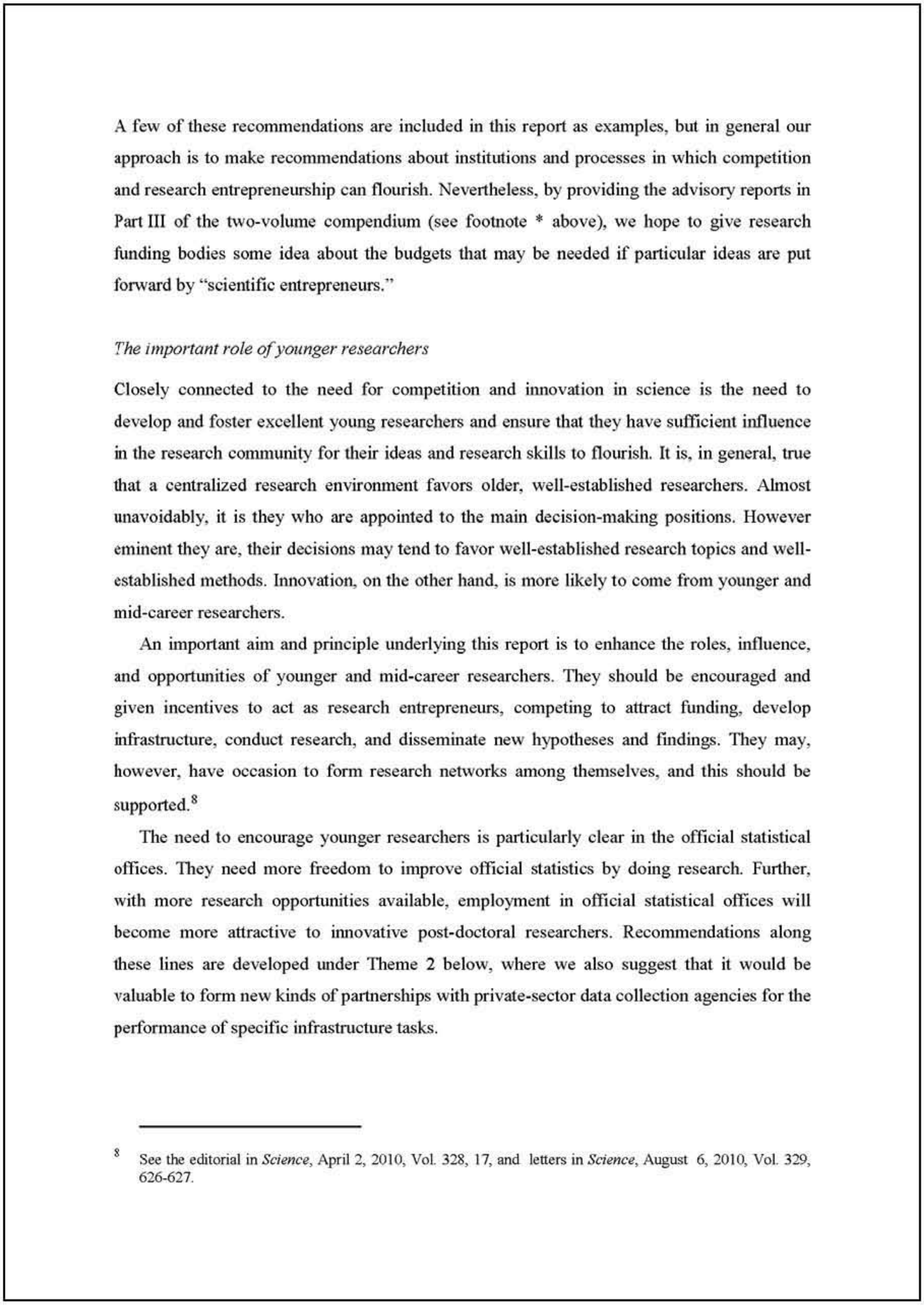
The history of Germany's Research Data Centers and Data Service Centers illustrates the same point. All the Research Data Centers and Data Service Centers established in the last six years were the result of independent initiatives intended to meet distinctive research needs. The KVI laid the groundwork by providing central funding for the first six Research Data Centers. But all the later ones basically relied on their own means, with minimal central funding. What was crucial was the basic concept for the Research Data Centers, and that was developed by the KVI in its 2001 report.

It is true that the German Data Forum (RatSWD) later institutionalized this framework by establishing a Standing Committee of the Research Data Centers and Data Service Centers (*Ständiger Ausschuss Forschungsdaten-Infrastruktur des RatSWD)*. This committee helps the centers to work together and put forward common interests, but it does not initiate new centers. Indeed, we believe that the committee should not do so. What is necessary is a common framework for new initiatives that aim to raise Germany's social science infrastructure to a higher level.

In this report we take some further steps towards developing a common framework for research infrastructure in the social sciences. In doing so, we bear in mind the increasing opportunities open to German researchers to contribute to European and international databases and projects, as well as to projects in Germany itself. We formulate some principles and highlight a range of concepts and ideas drawn from the advisory reports.[7]

We do not make detailed recommendations about specific research fields or particular infrastructural facilities. This would run counter to our view that innovative research directions and new ideas develop mainly at the grassroots of scientific and statistical communities. The advisory reports underlying these recommendations did include a large number of recommendations for promoting research in specific fields and on specific issues.

[7] The advisory reports are also summarized in the two-volume compendium– see Part II "Executive Summaries."



A few of these recommendations are included in this report as examples, but in general our approach is to make recommendations about institutions and processes in which competition and research entrepreneurship can flourish. Nevertheless, by providing the advisory reports in Part III of the two-volume compendium (see footnote * above), we hope to give research funding bodies some idea about the budgets that may be needed if particular ideas are put forward by "scientific entrepreneurs."

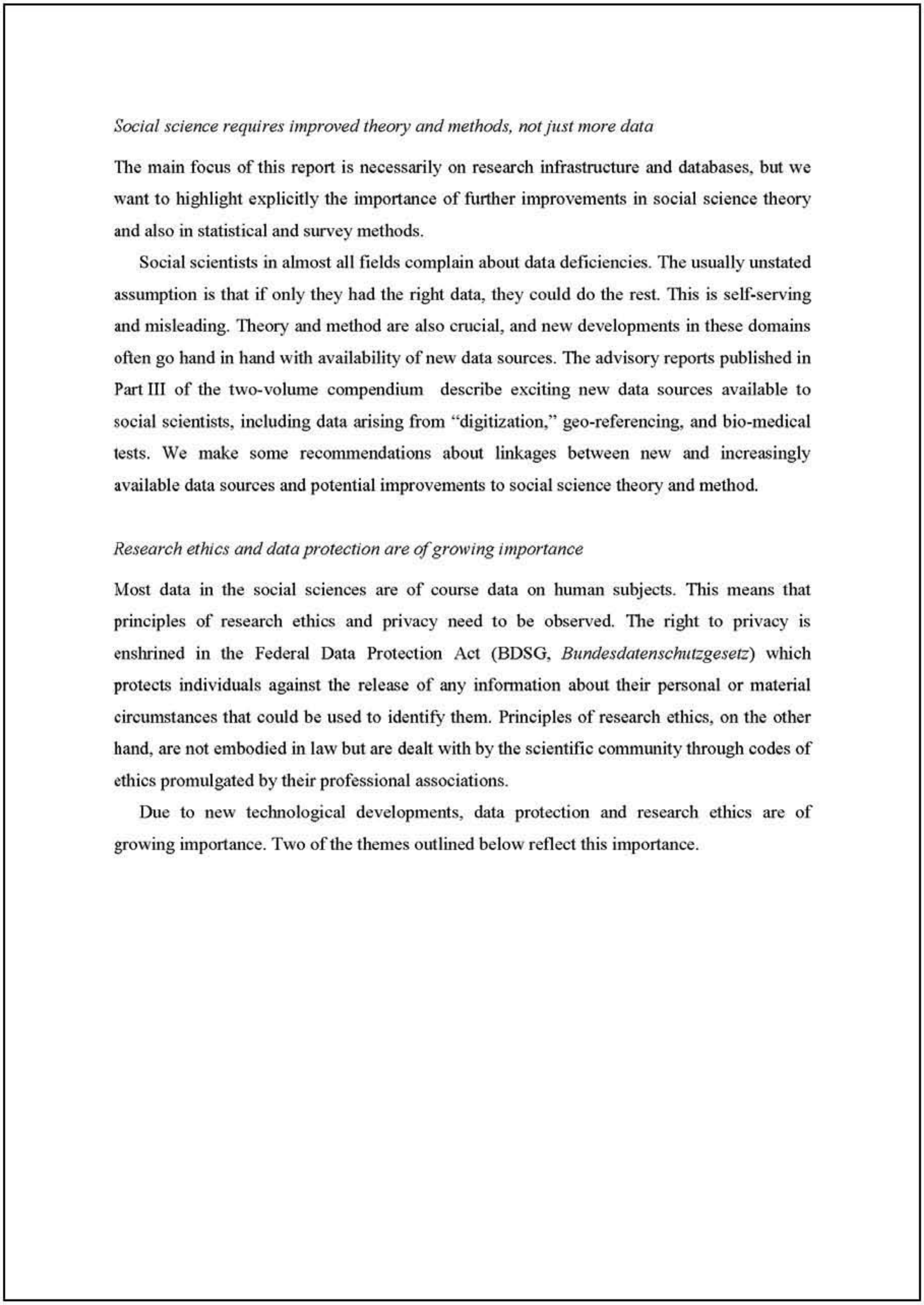
The important role of younger researchers

Closely connected to the need for competition and innovation in science is the need to develop and foster excellent young researchers and ensure that they have sufficient influence in the research community for their ideas and research skills to flourish. It is, in general, true that a centralized research environment favors older, well-established researchers. Almost unavoidably, it is they who are appointed to the main decision-making positions. However eminent they are, their decisions may tend to favor well-established research topics and well-established methods. Innovation, on the other hand, is more likely to come from younger and mid-career researchers.

An important aim and principle underlying this report is to enhance the roles, influence, and opportunities of younger and mid-career researchers. They should be encouraged and given incentives to act as research entrepreneurs, competing to attract funding, develop infrastructure, conduct research, and disseminate new hypotheses and findings. They may, however, have occasion to form research networks among themselves, and this should be supported.[8]

The need to encourage younger researchers is particularly clear in the official statistical offices. They need more freedom to improve official statistics by doing research. Further, with more research opportunities available, employment in official statistical offices will become more attractive to innovative post-doctoral researchers. Recommendations along these lines are developed under Theme 2 below, where we also suggest that it would be valuable to form new kinds of partnerships with private-sector data collection agencies for the performance of specific infrastructure tasks.

[8] See the editorial in *Science*, April 2, 2010, Vol. 328, 17, and letters in *Science*, August 6, 2010, Vol. 329, 626-627.



Social science requires improved theory and methods, not just more data

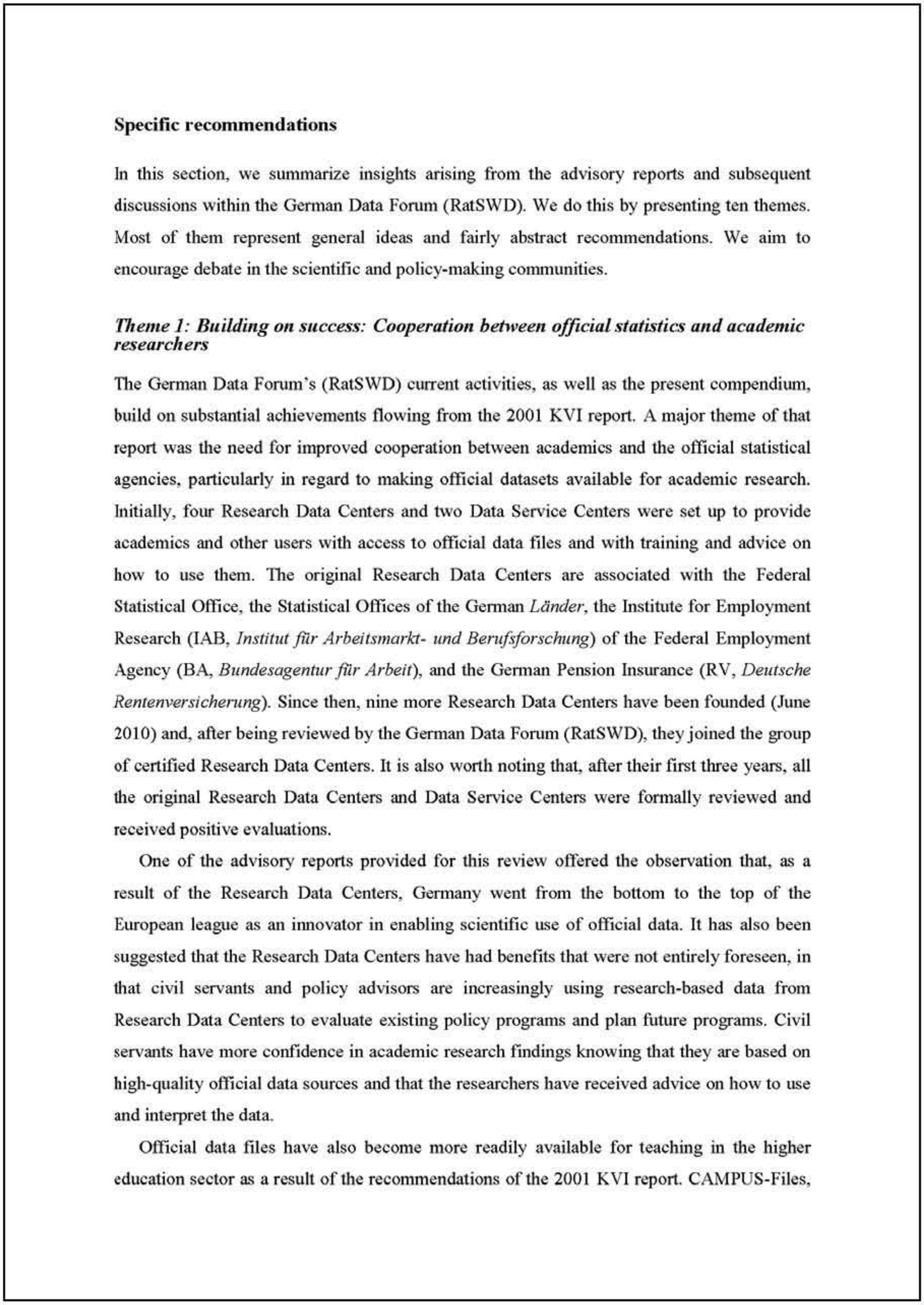
The main focus of this report is necessarily on research infrastructure and databases, but we want to highlight explicitly the importance of further improvements in social science theory and also in statistical and survey methods.

Social scientists in almost all fields complain about data deficiencies. The usually unstated assumption is that if only they had the right data, they could do the rest. This is self-serving and misleading. Theory and method are also crucial, and new developments in these domains often go hand in hand with availability of new data sources. The advisory reports published in Part III of the two-volume compendium describe exciting new data sources available to social scientists, including data arising from "digitization," geo-referencing, and bio-medical tests. We make some recommendations about linkages between new and increasingly available data sources and potential improvements to social science theory and method.

Research ethics and data protection are of growing importance

Most data in the social sciences are of course data on human subjects. This means that principles of research ethics and privacy need to be observed. The right to privacy is enshrined in the Federal Data Protection Act (BDSG, *Bundesdatenschutzgesetz*) which protects individuals against the release of any information about their personal or material circumstances that could be used to identify them. Principles of research ethics, on the other hand, are not embodied in law but are dealt with by the scientific community through codes of ethics promulgated by their professional associations.

Due to new technological developments, data protection and research ethics are of growing importance. Two of the themes outlined below reflect this importance.



Specific recommendations

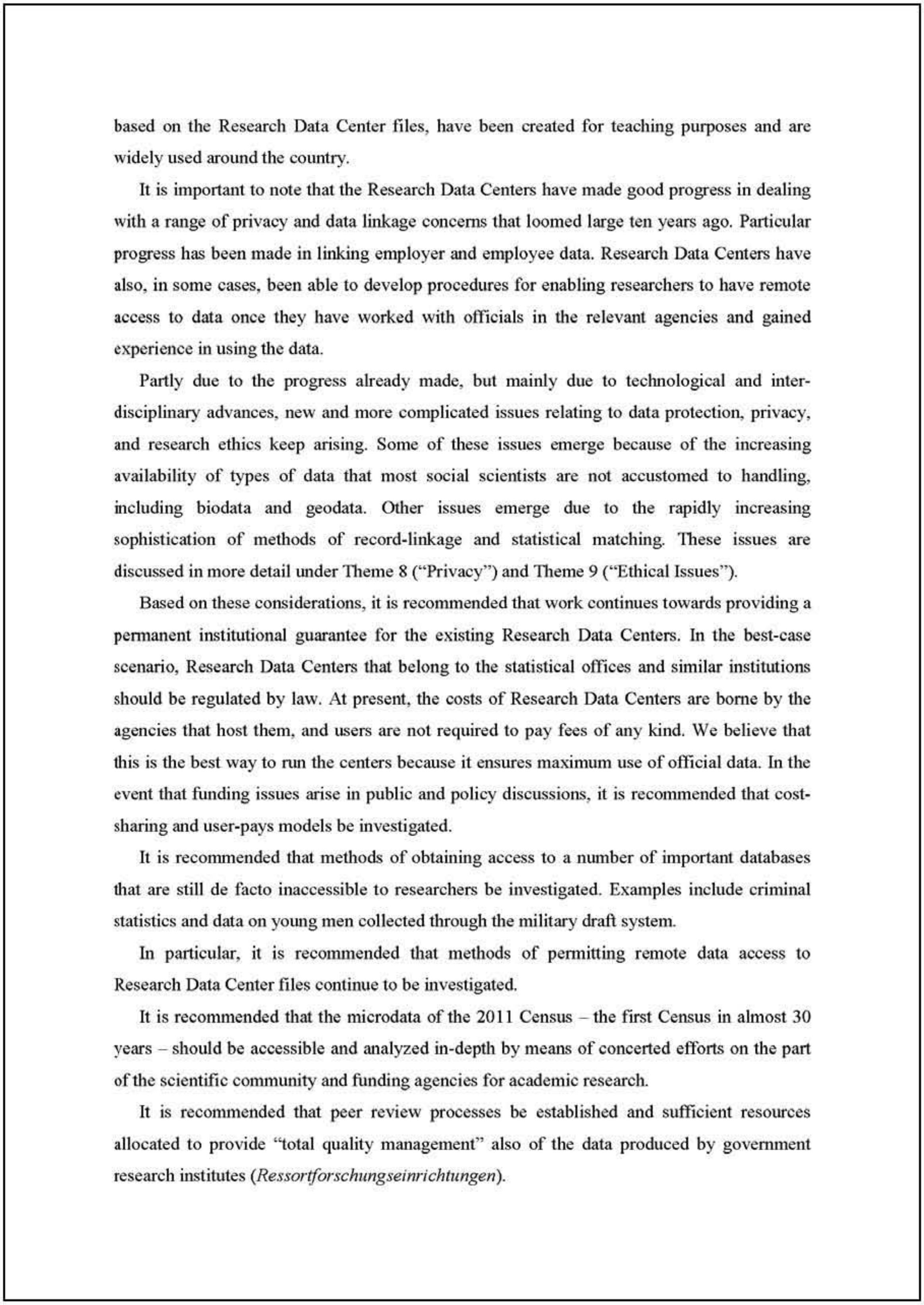
In this section, we summarize insights arising from the advisory reports and subsequent discussions within the German Data Forum (RatSWD). We do this by presenting ten themes. Most of them represent general ideas and fairly abstract recommendations. We aim to encourage debate in the scientific and policy-making communities.

Theme 1: Building on success: Cooperation between official statistics and academic researchers

The German Data Forum's (RatSWD) current activities, as well as the present compendium, build on substantial achievements flowing from the 2001 KVI report. A major theme of that report was the need for improved cooperation between academics and the official statistical agencies, particularly in regard to making official datasets available for academic research. Initially, four Research Data Centers and two Data Service Centers were set up to provide academics and other users with access to official data files and with training and advice on how to use them. The original Research Data Centers are associated with the Federal Statistical Office, the Statistical Offices of the German *Länder*, the Institute for Employment Research (IAB, *Institut für Arbeitsmarkt- und Berufsforschung*) of the Federal Employment Agency (BA, *Bundesagentur für Arbeit*), and the German Pension Insurance (RV, *Deutsche Rentenversicherung*). Since then, nine more Research Data Centers have been founded (June 2010) and, after being reviewed by the German Data Forum (RatSWD), they joined the group of certified Research Data Centers. It is also worth noting that, after their first three years, all the original Research Data Centers and Data Service Centers were formally reviewed and received positive evaluations.

One of the advisory reports provided for this review offered the observation that, as a result of the Research Data Centers, Germany went from the bottom to the top of the European league as an innovator in enabling scientific use of official data. It has also been suggested that the Research Data Centers have had benefits that were not entirely foreseen, in that civil servants and policy advisors are increasingly using research-based data from Research Data Centers to evaluate existing policy programs and plan future programs. Civil servants have more confidence in academic research findings knowing that they are based on high-quality official data sources and that the researchers have received advice on how to use and interpret the data.

Official data files have also become more readily available for teaching in the higher education sector as a result of the recommendations of the 2001 KVI report. CAMPUS-Files,



based on the Research Data Center files, have been created for teaching purposes and are widely used around the country.

It is important to note that the Research Data Centers have made good progress in dealing with a range of privacy and data linkage concerns that loomed large ten years ago. Particular progress has been made in linking employer and employee data. Research Data Centers have also, in some cases, been able to develop procedures for enabling researchers to have remote access to data once they have worked with officials in the relevant agencies and gained experience in using the data.

Partly due to the progress already made, but mainly due to technological and interdisciplinary advances, new and more complicated issues relating to data protection, privacy, and research ethics keep arising. Some of these issues emerge because of the increasing availability of types of data that most social scientists are not accustomed to handling, including biodata and geodata. Other issues emerge due to the rapidly increasing sophistication of methods of record-linkage and statistical matching. These issues are discussed in more detail under Theme 8 ("Privacy") and Theme 9 ("Ethical Issues").

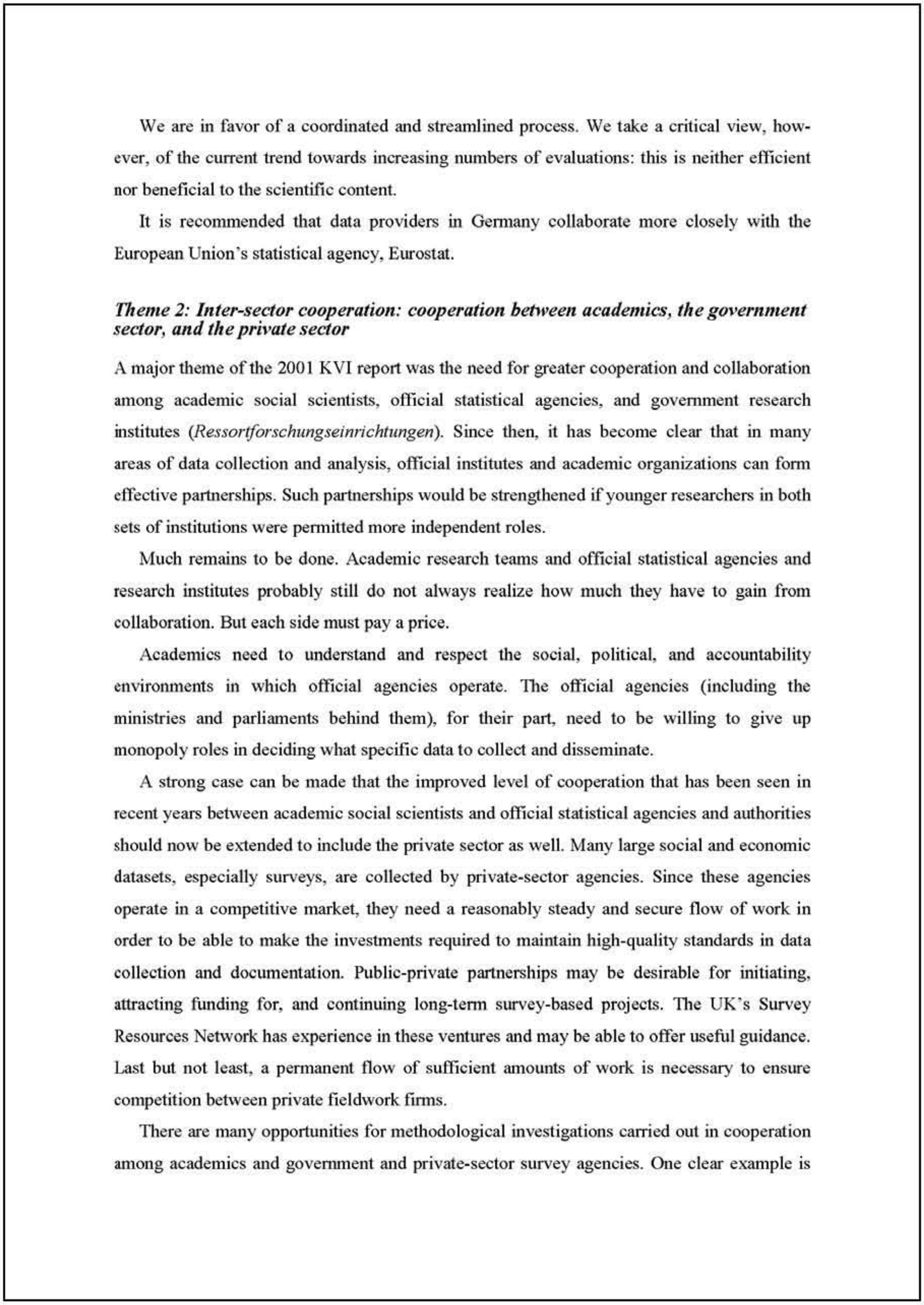
Based on these considerations, it is recommended that work continues towards providing a permanent institutional guarantee for the existing Research Data Centers. In the best-case scenario, Research Data Centers that belong to the statistical offices and similar institutions should be regulated by law. At present, the costs of Research Data Centers are borne by the agencies that host them, and users are not required to pay fees of any kind. We believe that this is the best way to run the centers because it ensures maximum use of official data. In the event that funding issues arise in public and policy discussions, it is recommended that cost-sharing and user-pays models be investigated.

It is recommended that methods of obtaining access to a number of important databases that are still de facto inaccessible to researchers be investigated. Examples include criminal statistics and data on young men collected through the military draft system.

In particular, it is recommended that methods of permitting remote data access to Research Data Center files continue to be investigated.

It is recommended that the microdata of the 2011 Census – the first Census in almost 30 years – should be accessible and analyzed in-depth by means of concerted efforts on the part of the scientific community and funding agencies for academic research.

It is recommended that peer review processes be established and sufficient resources allocated to provide "total quality management" also of the data produced by government research institutes (*Ressortforschungseinrichtungen*).



We are in favor of a coordinated and streamlined process. We take a critical view, however, of the current trend towards increasing numbers of evaluations: this is neither efficient nor beneficial to the scientific content.

It is recommended that data providers in Germany collaborate more closely with the European Union's statistical agency, Eurostat.

Theme 2: Inter-sector cooperation: cooperation between academics, the government sector, and the private sector

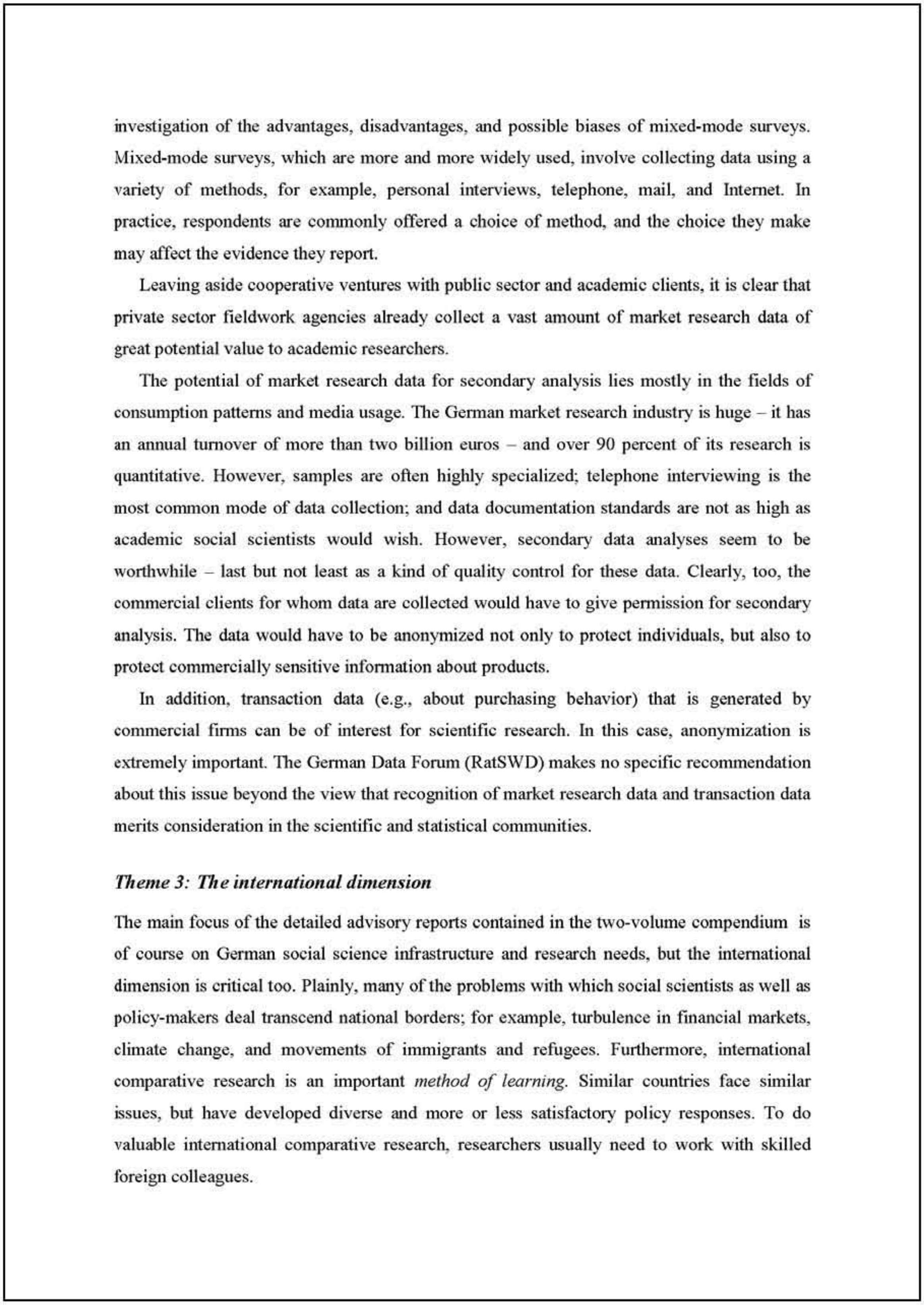
A major theme of the 2001 KVI report was the need for greater cooperation and collaboration among academic social scientists, official statistical agencies, and government research institutes (*Ressortforschungseinrichtungen*). Since then, it has become clear that in many areas of data collection and analysis, official institutes and academic organizations can form effective partnerships. Such partnerships would be strengthened if younger researchers in both sets of institutions were permitted more independent roles.

Much remains to be done. Academic research teams and official statistical agencies and research institutes probably still do not always realize how much they have to gain from collaboration. But each side must pay a price.

Academics need to understand and respect the social, political, and accountability environments in which official agencies operate. The official agencies (including the ministries and parliaments behind them), for their part, need to be willing to give up monopoly roles in deciding what specific data to collect and disseminate.

A strong case can be made that the improved level of cooperation that has been seen in recent years between academic social scientists and official statistical agencies and authorities should now be extended to include the private sector as well. Many large social and economic datasets, especially surveys, are collected by private-sector agencies. Since these agencies operate in a competitive market, they need a reasonably steady and secure flow of work in order to be able to make the investments required to maintain high-quality standards in data collection and documentation. Public-private partnerships may be desirable for initiating, attracting funding for, and continuing long-term survey-based projects. The UK's Survey Resources Network has experience in these ventures and may be able to offer useful guidance. Last but not least, a permanent flow of sufficient amounts of work is necessary to ensure competition between private fieldwork firms.

There are many opportunities for methodological investigations carried out in cooperation among academics and government and private-sector survey agencies. One clear example is



investigation of the advantages, disadvantages, and possible biases of mixed-mode surveys. Mixed-mode surveys, which are more and more widely used, involve collecting data using a variety of methods, for example, personal interviews, telephone, mail, and Internet. In practice, respondents are commonly offered a choice of method, and the choice they make may affect the evidence they report.

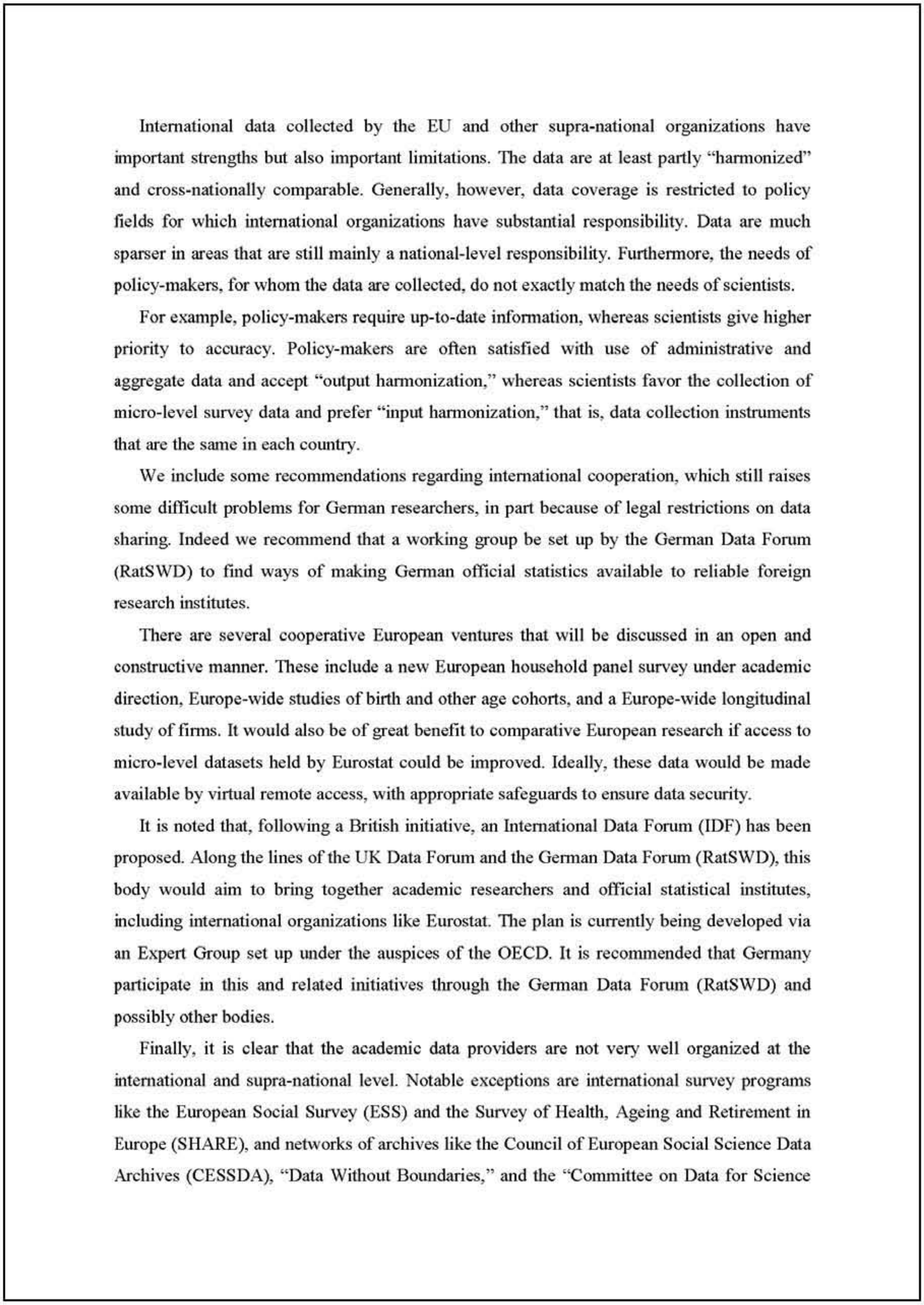
Leaving aside cooperative ventures with public sector and academic clients, it is clear that private sector fieldwork agencies already collect a vast amount of market research data of great potential value to academic researchers.

The potential of market research data for secondary analysis lies mostly in the fields of consumption patterns and media usage. The German market research industry is huge – it has an annual turnover of more than two billion euros – and over 90 percent of its research is quantitative. However, samples are often highly specialized; telephone interviewing is the most common mode of data collection; and data documentation standards are not as high as academic social scientists would wish. However, secondary data analyses seem to be worthwhile – last but not least as a kind of quality control for these data. Clearly, too, the commercial clients for whom data are collected would have to give permission for secondary analysis. The data would have to be anonymized not only to protect individuals, but also to protect commercially sensitive information about products.

In addition, transaction data (e.g., about purchasing behavior) that is generated by commercial firms can be of interest for scientific research. In this case, anonymization is extremely important. The German Data Forum (RatSWD) makes no specific recommendation about this issue beyond the view that recognition of market research data and transaction data merits consideration in the scientific and statistical communities.

Theme 3: The international dimension

The main focus of the detailed advisory reports contained in the two-volume compendium is of course on German social science infrastructure and research needs, but the international dimension is critical too. Plainly, many of the problems with which social scientists as well as policy-makers deal transcend national borders; for example, turbulence in financial markets, climate change, and movements of immigrants and refugees. Furthermore, international comparative research is an important *method of learning*. Similar countries face similar issues, but have developed diverse and more or less satisfactory policy responses. To do valuable international comparative research, researchers usually need to work with skilled foreign colleagues.



International data collected by the EU and other supra-national organizations have important strengths but also important limitations. The data are at least partly "harmonized" and cross-nationally comparable. Generally, however, data coverage is restricted to policy fields for which international organizations have substantial responsibility. Data are much sparser in areas that are still mainly a national-level responsibility. Furthermore, the needs of policy-makers, for whom the data are collected, do not exactly match the needs of scientists.

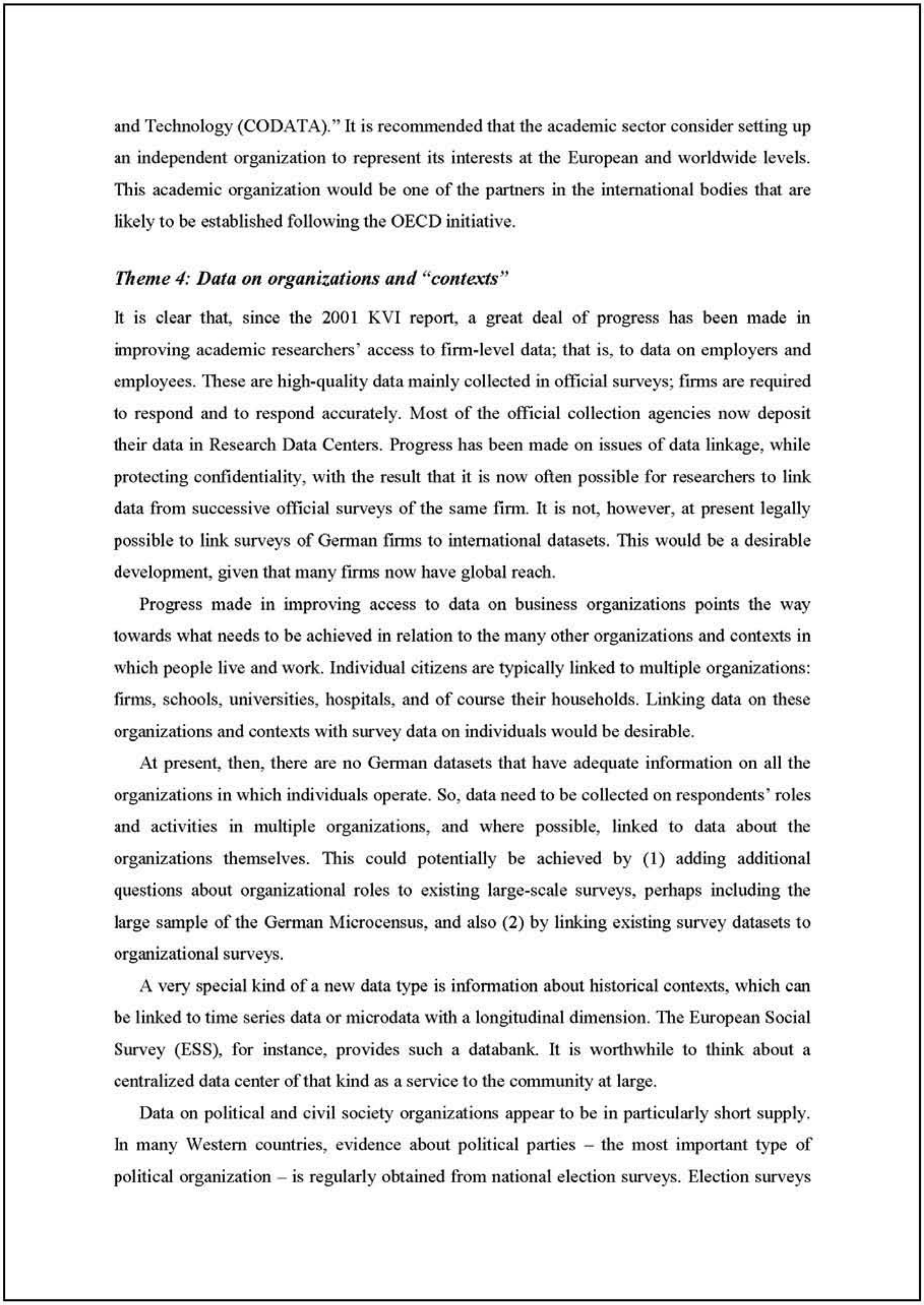
For example, policy-makers require up-to-date information, whereas scientists give higher priority to accuracy. Policy-makers are often satisfied with use of administrative and aggregate data and accept "output harmonization," whereas scientists favor the collection of micro-level survey data and prefer "input harmonization," that is, data collection instruments that are the same in each country.

We include some recommendations regarding international cooperation, which still raises some difficult problems for German researchers, in part because of legal restrictions on data sharing. Indeed we recommend that a working group be set up by the German Data Forum (RatSWD) to find ways of making German official statistics available to reliable foreign research institutes.

There are several cooperative European ventures that will be discussed in an open and constructive manner. These include a new European household panel survey under academic direction, Europe-wide studies of birth and other age cohorts, and a Europe-wide longitudinal study of firms. It would also be of great benefit to comparative European research if access to micro-level datasets held by Eurostat could be improved. Ideally, these data would be made available by virtual remote access, with appropriate safeguards to ensure data security.

It is noted that, following a British initiative, an International Data Forum (IDF) has been proposed. Along the lines of the UK Data Forum and the German Data Forum (RatSWD), this body would aim to bring together academic researchers and official statistical institutes, including international organizations like Eurostat. The plan is currently being developed via an Expert Group set up under the auspices of the OECD. It is recommended that Germany participate in this and related initiatives through the German Data Forum (RatSWD) and possibly other bodies.

Finally, it is clear that the academic data providers are not very well organized at the international and supra-national level. Notable exceptions are international survey programs like the European Social Survey (ESS) and the Survey of Health, Ageing and Retirement in Europe (SHARE), and networks of archives like the Council of European Social Science Data Archives (CESSDA), "Data Without Boundaries," and the "Committee on Data for Science



and Technology (CODATA)." It is recommended that the academic sector consider setting up an independent organization to represent its interests at the European and worldwide levels. This academic organization would be one of the partners in the international bodies that are likely to be established following the OECD initiative.

Theme 4: Data on organizations and "contexts"

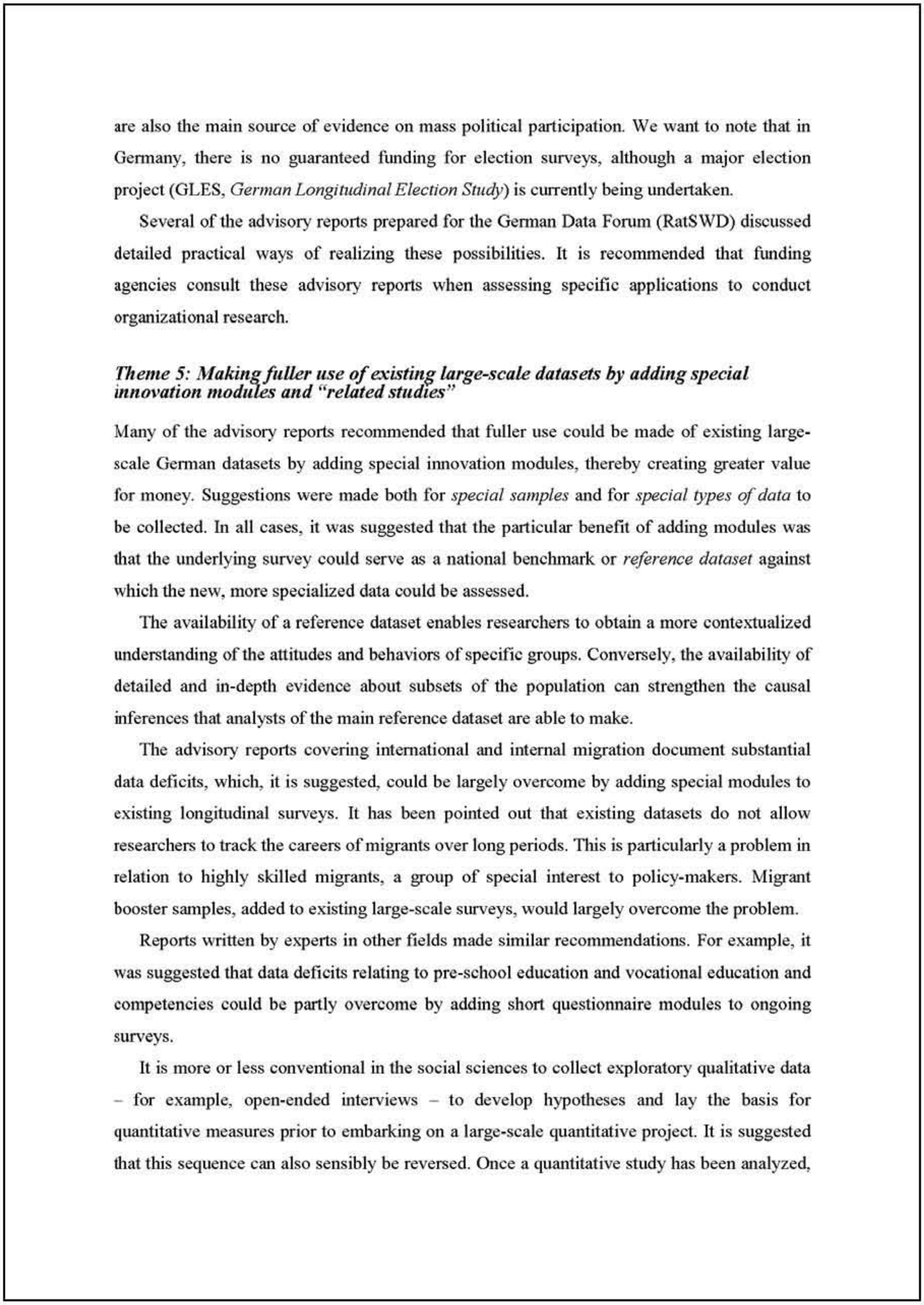
It is clear that, since the 2001 KVI report, a great deal of progress has been made in improving academic researchers' access to firm-level data; that is, to data on employers and employees. These are high-quality data mainly collected in official surveys; firms are required to respond and to respond accurately. Most of the official collection agencies now deposit their data in Research Data Centers. Progress has been made on issues of data linkage, while protecting confidentiality, with the result that it is now often possible for researchers to link data from successive official surveys of the same firm. It is not, however, at present legally possible to link surveys of German firms to international datasets. This would be a desirable development, given that many firms now have global reach.

Progress made in improving access to data on business organizations points the way towards what needs to be achieved in relation to the many other organizations and contexts in which people live and work. Individual citizens are typically linked to multiple organizations: firms, schools, universities, hospitals, and of course their households. Linking data on these organizations and contexts with survey data on individuals would be desirable.

At present, then, there are no German datasets that have adequate information on all the organizations in which individuals operate. So, data need to be collected on respondents' roles and activities in multiple organizations, and where possible, linked to data about the organizations themselves. This could potentially be achieved by (1) adding additional questions about organizational roles to existing large-scale surveys, perhaps including the large sample of the German Microcensus, and also (2) by linking existing survey datasets to organizational surveys.

A very special kind of a new data type is information about historical contexts, which can be linked to time series data or microdata with a longitudinal dimension. The European Social Survey (ESS), for instance, provides such a databank. It is worthwhile to think about a centralized data center of that kind as a service to the community at large.

Data on political and civil society organizations appear to be in particularly short supply. In many Western countries, evidence about political parties – the most important type of political organization – is regularly obtained from national election surveys. Election surveys



are also the main source of evidence on mass political participation. We want to note that in Germany, there is no guaranteed funding for election surveys, although a major election project (GLES, *German Longitudinal Election Study*) is currently being undertaken.

Several of the advisory reports prepared for the German Data Forum (RatSWD) discussed detailed practical ways of realizing these possibilities. It is recommended that funding agencies consult these advisory reports when assessing specific applications to conduct organizational research.

Theme 5: Making fuller use of existing large-scale datasets by adding special innovation modules and "related studies"

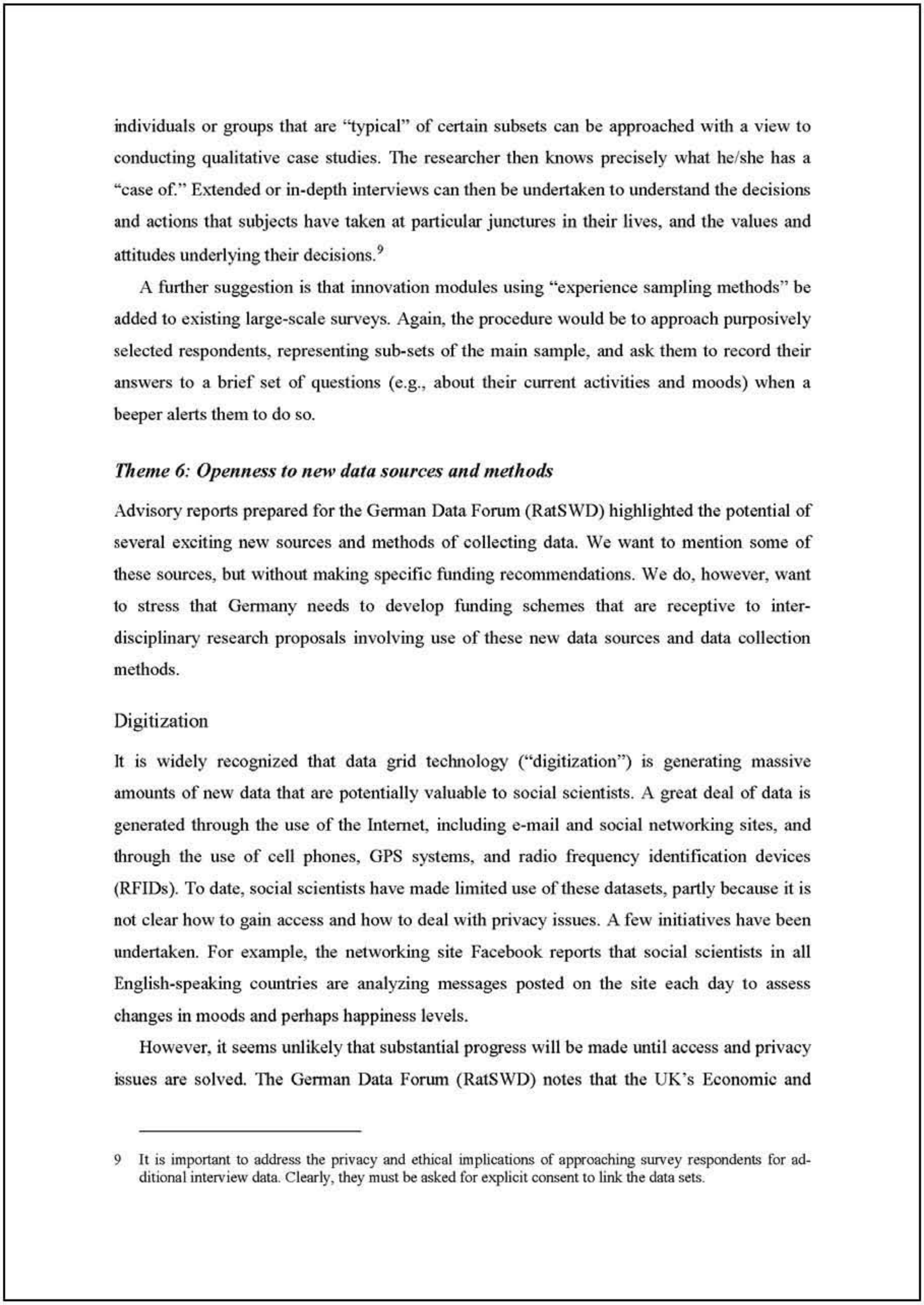
Many of the advisory reports recommended that fuller use could be made of existing large-scale German datasets by adding special innovation modules, thereby creating greater value for money. Suggestions were made both for *special samples* and for *special types of data* to be collected. In all cases, it was suggested that the particular benefit of adding modules was that the underlying survey could serve as a national benchmark or *reference dataset* against which the new, more specialized data could be assessed.

The availability of a reference dataset enables researchers to obtain a more contextualized understanding of the attitudes and behaviors of specific groups. Conversely, the availability of detailed and in-depth evidence about subsets of the population can strengthen the causal inferences that analysts of the main reference dataset are able to make.

The advisory reports covering international and internal migration document substantial data deficits, which, it is suggested, could be largely overcome by adding special modules to existing longitudinal surveys. It has been pointed out that existing datasets do not allow researchers to track the careers of migrants over long periods. This is particularly a problem in relation to highly skilled migrants, a group of special interest to policy-makers. Migrant booster samples, added to existing large-scale surveys, would largely overcome the problem.

Reports written by experts in other fields made similar recommendations. For example, it was suggested that data deficits relating to pre-school education and vocational education and competencies could be partly overcome by adding short questionnaire modules to ongoing surveys.

It is more or less conventional in the social sciences to collect exploratory qualitative data – for example, open-ended interviews – to develop hypotheses and lay the basis for quantitative measures prior to embarking on a large-scale quantitative project. It is suggested that this sequence can also sensibly be reversed. Once a quantitative study has been analyzed,



individuals or groups that are "typical" of certain subsets can be approached with a view to conducting qualitative case studies. The researcher then knows precisely what he/she has a "case of." Extended or in-depth interviews can then be undertaken to understand the decisions and actions that subjects have taken at particular junctures in their lives, and the values and attitudes underlying their decisions.[9]

A further suggestion is that innovation modules using "experience sampling methods" be added to existing large-scale surveys. Again, the procedure would be to approach purposively selected respondents, representing sub-sets of the main sample, and ask them to record their answers to a brief set of questions (e.g., about their current activities and moods) when a beeper alerts them to do so.

Theme 6: Openness to new data sources and methods

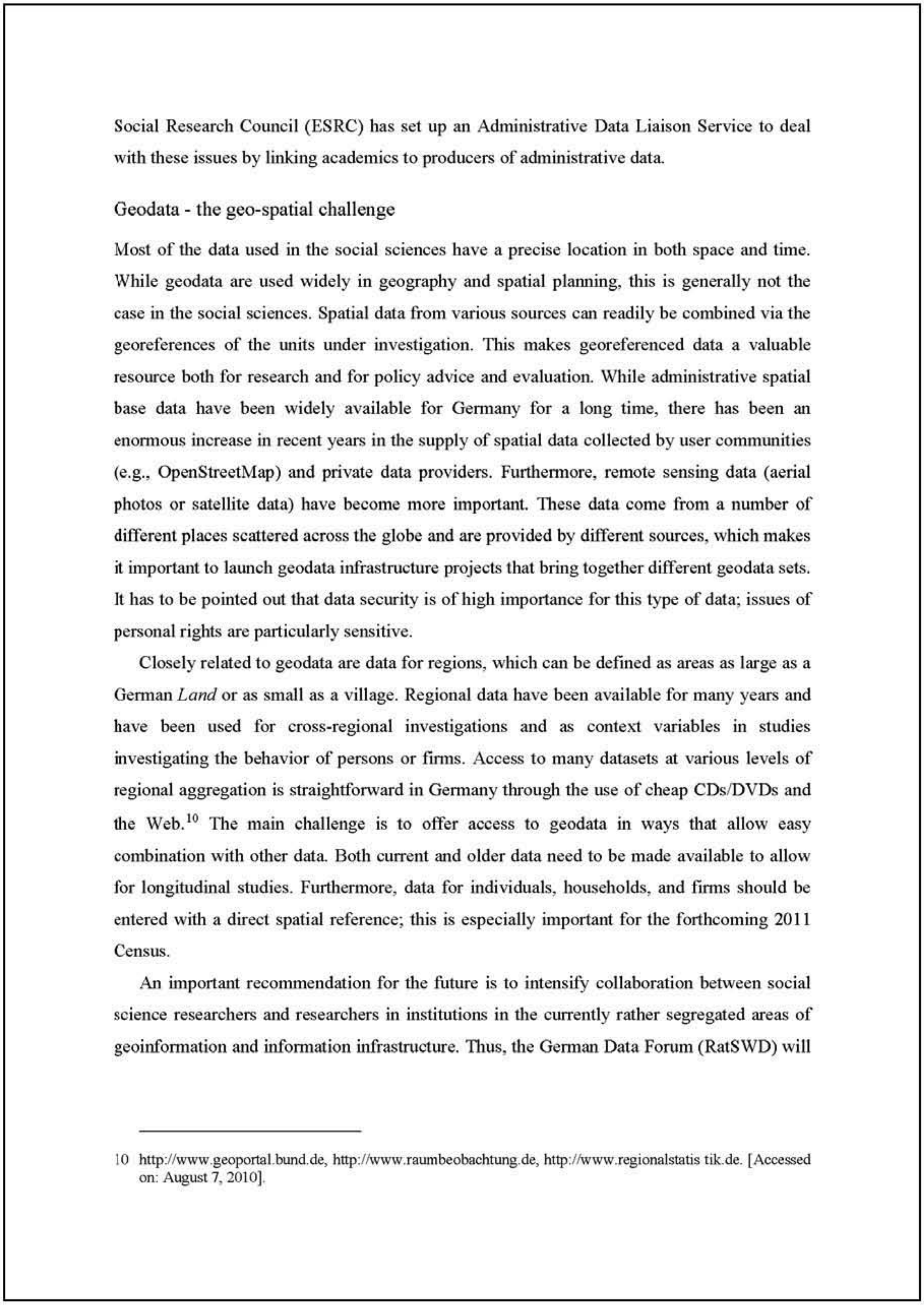
Advisory reports prepared for the German Data Forum (RatSWD) highlighted the potential of several exciting new sources and methods of collecting data. We want to mention some of these sources, but without making specific funding recommendations. We do, however, want to stress that Germany needs to develop funding schemes that are receptive to interdisciplinary research proposals involving use of these new data sources and data collection methods.

Digitization

It is widely recognized that data grid technology ("digitization") is generating massive amounts of new data that are potentially valuable to social scientists. A great deal of data is generated through the use of the Internet, including e-mail and social networking sites, and through the use of cell phones, GPS systems, and radio frequency identification devices (RFIDs). To date, social scientists have made limited use of these datasets, partly because it is not clear how to gain access and how to deal with privacy issues. A few initiatives have been undertaken. For example, the networking site Facebook reports that social scientists in all English-speaking countries are analyzing messages posted on the site each day to assess changes in moods and perhaps happiness levels.

However, it seems unlikely that substantial progress will be made until access and privacy issues are solved. The German Data Forum (RatSWD) notes that the UK's Economic and

[9] It is important to address the privacy and ethical implications of approaching survey respondents for additional interview data. Clearly, they must be asked for explicit consent to link the data sets.



Social Research Council (ESRC) has set up an Administrative Data Liaison Service to deal with these issues by linking academics to producers of administrative data.

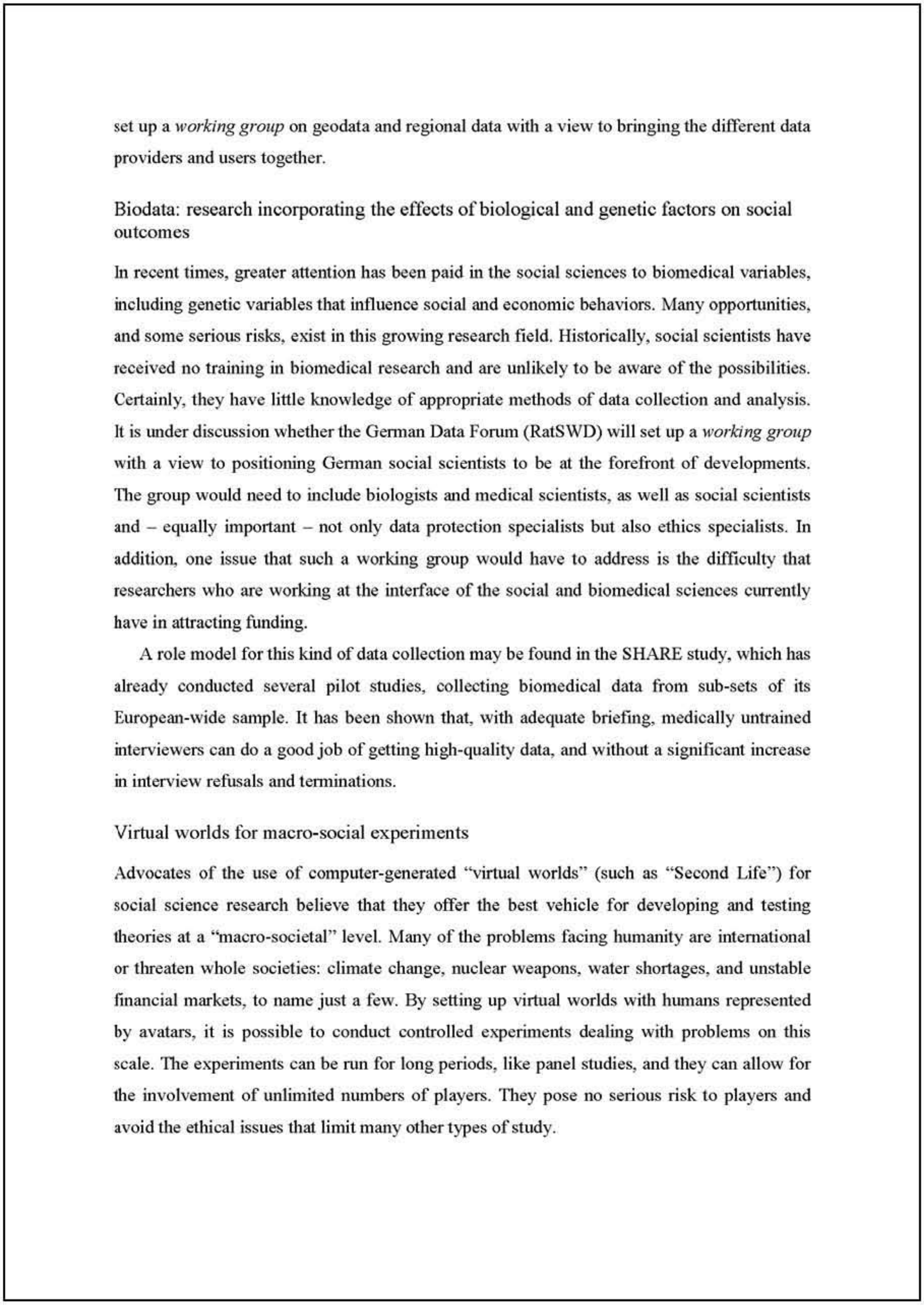
Geodata - the geo-spatial challenge

Most of the data used in the social sciences have a precise location in both space and time. While geodata are used widely in geography and spatial planning, this is generally not the case in the social sciences. Spatial data from various sources can readily be combined via the georeferences of the units under investigation. This makes georeferenced data a valuable resource both for research and for policy advice and evaluation. While administrative spatial base data have been widely available for Germany for a long time, there has been an enormous increase in recent years in the supply of spatial data collected by user communities (e.g., OpenStreetMap) and private data providers. Furthermore, remote sensing data (aerial photos or satellite data) have become more important. These data come from a number of different places scattered across the globe and are provided by different sources, which makes it important to launch geodata infrastructure projects that bring together different geodata sets. It has to be pointed out that data security is of high importance for this type of data; issues of personal rights are particularly sensitive.

Closely related to geodata are data for regions, which can be defined as areas as large as a German *Land* or as small as a village. Regional data have been available for many years and have been used for cross-regional investigations and as context variables in studies investigating the behavior of persons or firms. Access to many datasets at various levels of regional aggregation is straightforward in Germany through the use of cheap CDs/DVDs and the Web.[10] The main challenge is to offer access to geodata in ways that allow easy combination with other data. Both current and older data need to be made available to allow for longitudinal studies. Furthermore, data for individuals, households, and firms should be entered with a direct spatial reference; this is especially important for the forthcoming 2011 Census.

An important recommendation for the future is to intensify collaboration between social science researchers and researchers in institutions in the currently rather segregated areas of geoinformation and information infrastructure. Thus, the German Data Forum (RatSWD) will

[10] http://www.geoportal.bund.de, http://www.raumbeobachtung.de, http://www.regionalstatis tik.de. [Accessed on: August 7, 2010].



set up a *working group* on geodata and regional data with a view to bringing the different data providers and users together.

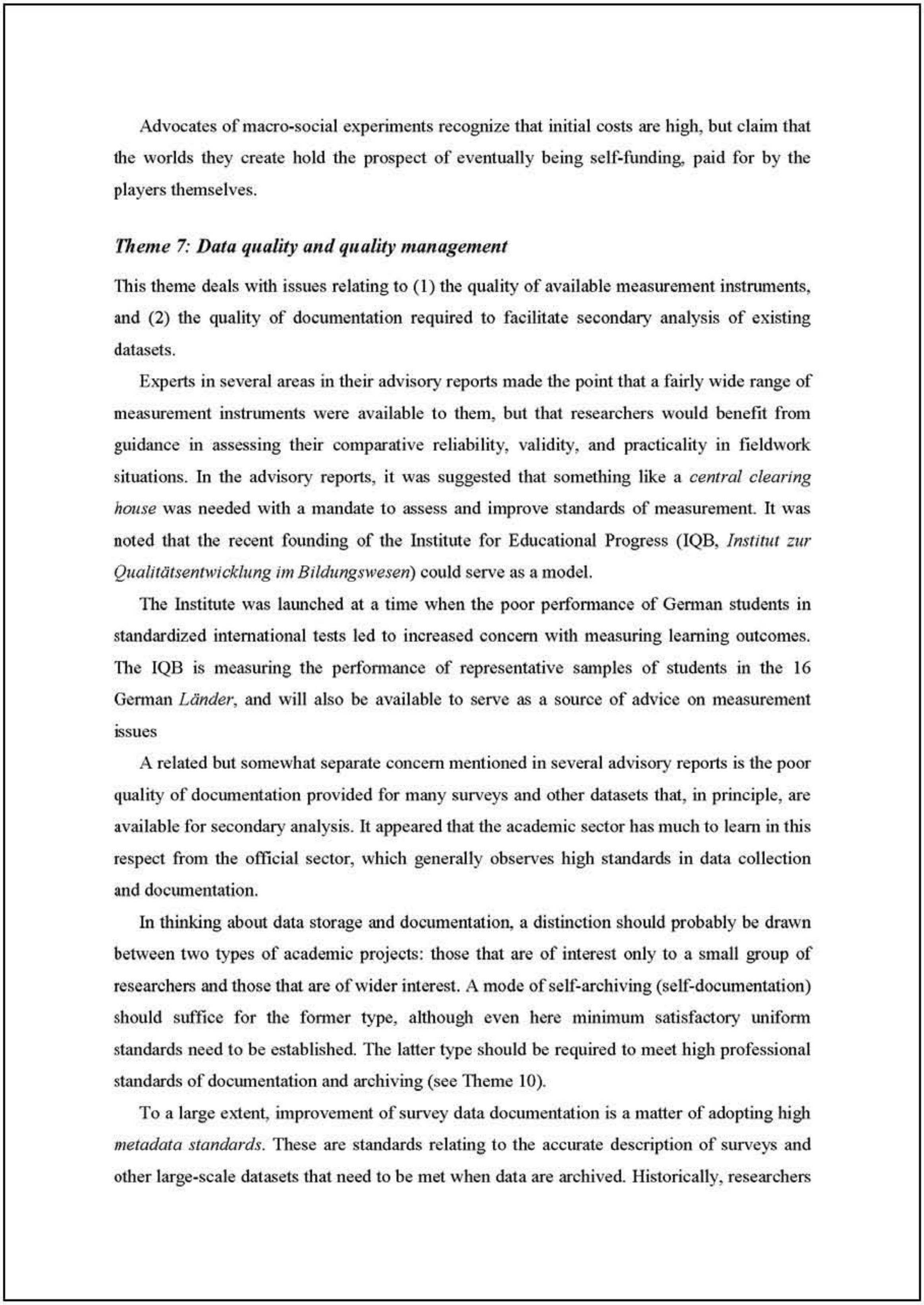
Biodata: research incorporating the effects of biological and genetic factors on social outcomes

In recent times, greater attention has been paid in the social sciences to biomedical variables, including genetic variables that influence social and economic behaviors. Many opportunities, and some serious risks, exist in this growing research field. Historically, social scientists have received no training in biomedical research and are unlikely to be aware of the possibilities. Certainly, they have little knowledge of appropriate methods of data collection and analysis. It is under discussion whether the German Data Forum (RatSWD) will set up a *working group* with a view to positioning German social scientists to be at the forefront of developments. The group would need to include biologists and medical scientists, as well as social scientists and – equally important – not only data protection specialists but also ethics specialists. In addition, one issue that such a working group would have to address is the difficulty that researchers who are working at the interface of the social and biomedical sciences currently have in attracting funding.

A role model for this kind of data collection may be found in the SHARE study, which has already conducted several pilot studies, collecting biomedical data from sub-sets of its European-wide sample. It has been shown that, with adequate briefing, medically untrained interviewers can do a good job of getting high-quality data, and without a significant increase in interview refusals and terminations.

Virtual worlds for macro-social experiments

Advocates of the use of computer-generated "virtual worlds" (such as "Second Life") for social science research believe that they offer the best vehicle for developing and testing theories at a "macro-societal" level. Many of the problems facing humanity are international or threaten whole societies: climate change, nuclear weapons, water shortages, and unstable financial markets, to name just a few. By setting up virtual worlds with humans represented by avatars, it is possible to conduct controlled experiments dealing with problems on this scale. The experiments can be run for long periods, like panel studies, and they can allow for the involvement of unlimited numbers of players. They pose no serious risk to players and avoid the ethical issues that limit many other types of study.



Advocates of macro-social experiments recognize that initial costs are high, but claim that the worlds they create hold the prospect of eventually being self-funding, paid for by the players themselves.

Theme 7: Data quality and quality management

This theme deals with issues relating to (1) the quality of available measurement instruments, and (2) the quality of documentation required to facilitate secondary analysis of existing datasets.

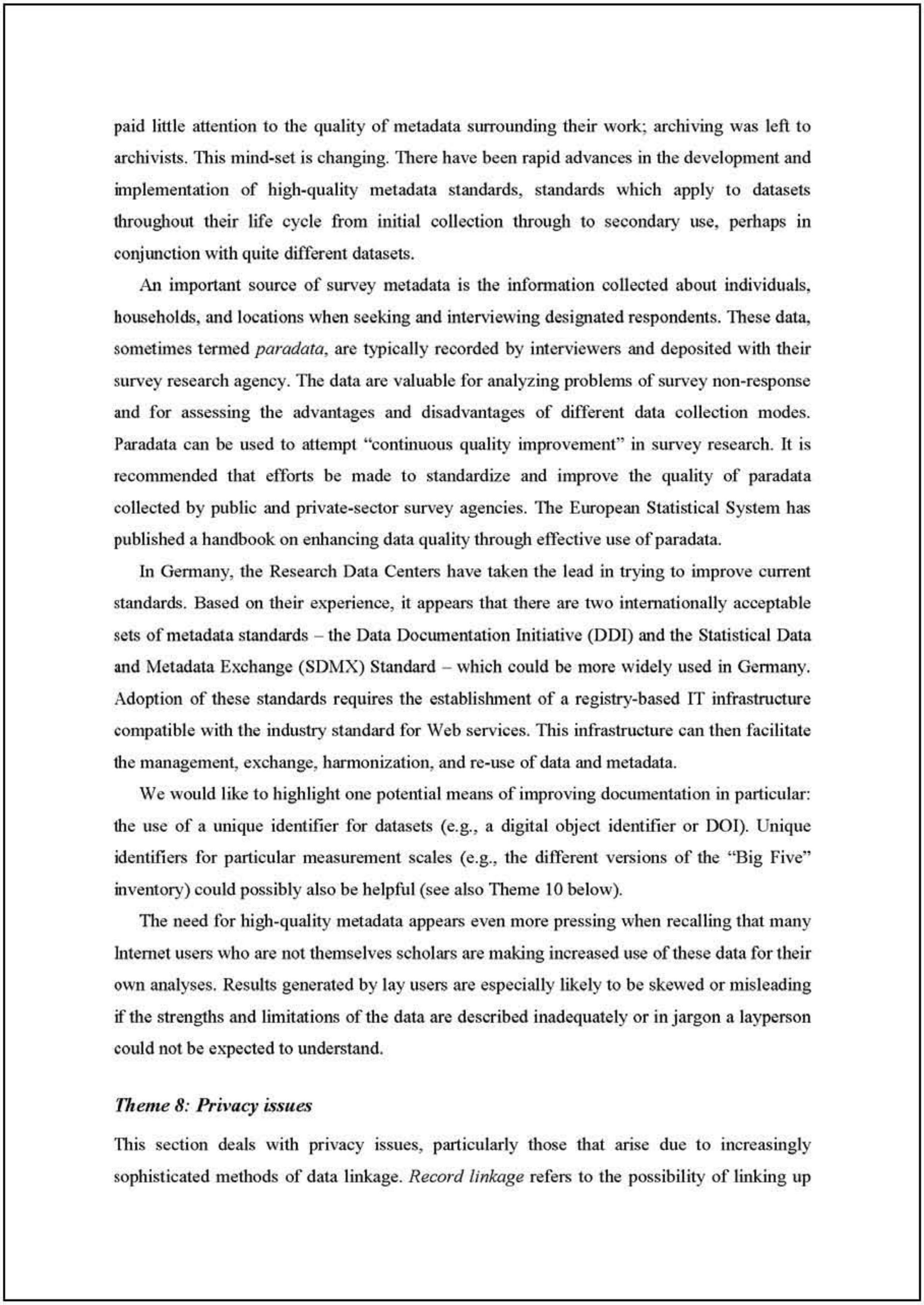
Experts in several areas in their advisory reports made the point that a fairly wide range of measurement instruments were available to them, but that researchers would benefit from guidance in assessing their comparative reliability, validity, and practicality in fieldwork situations. In the advisory reports, it was suggested that something like a *central clearing house* was needed with a mandate to assess and improve standards of measurement. It was noted that the recent founding of the Institute for Educational Progress (IQB, *Institut zur Qualitätsentwicklung im Bildungswesen*) could serve as a model.

The Institute was launched at a time when the poor performance of German students in standardized international tests led to increased concern with measuring learning outcomes. The IQB is measuring the performance of representative samples of students in the 16 German *Länder*, and will also be available to serve as a source of advice on measurement issues

A related but somewhat separate concern mentioned in several advisory reports is the poor quality of documentation provided for many surveys and other datasets that, in principle, are available for secondary analysis. It appeared that the academic sector has much to learn in this respect from the official sector, which generally observes high standards in data collection and documentation.

In thinking about data storage and documentation, a distinction should probably be drawn between two types of academic projects: those that are of interest only to a small group of researchers and those that are of wider interest. A mode of self-archiving (self-documentation) should suffice for the former type, although even here minimum satisfactory uniform standards need to be established. The latter type should be required to meet high professional standards of documentation and archiving (see Theme 10).

To a large extent, improvement of survey data documentation is a matter of adopting high *metadata standards*. These are standards relating to the accurate description of surveys and other large-scale datasets that need to be met when data are archived. Historically, researchers



paid little attention to the quality of metadata surrounding their work; archiving was left to archivists. This mind-set is changing. There have been rapid advances in the development and implementation of high-quality metadata standards, standards which apply to datasets throughout their life cycle from initial collection through to secondary use, perhaps in conjunction with quite different datasets.

An important source of survey metadata is the information collected about individuals, households, and locations when seeking and interviewing designated respondents. These data, sometimes termed *paradata*, are typically recorded by interviewers and deposited with their survey research agency. The data are valuable for analyzing problems of survey non-response and for assessing the advantages and disadvantages of different data collection modes. Paradata can be used to attempt "continuous quality improvement" in survey research. It is recommended that efforts be made to standardize and improve the quality of paradata collected by public and private-sector survey agencies. The European Statistical System has published a handbook on enhancing data quality through effective use of paradata.

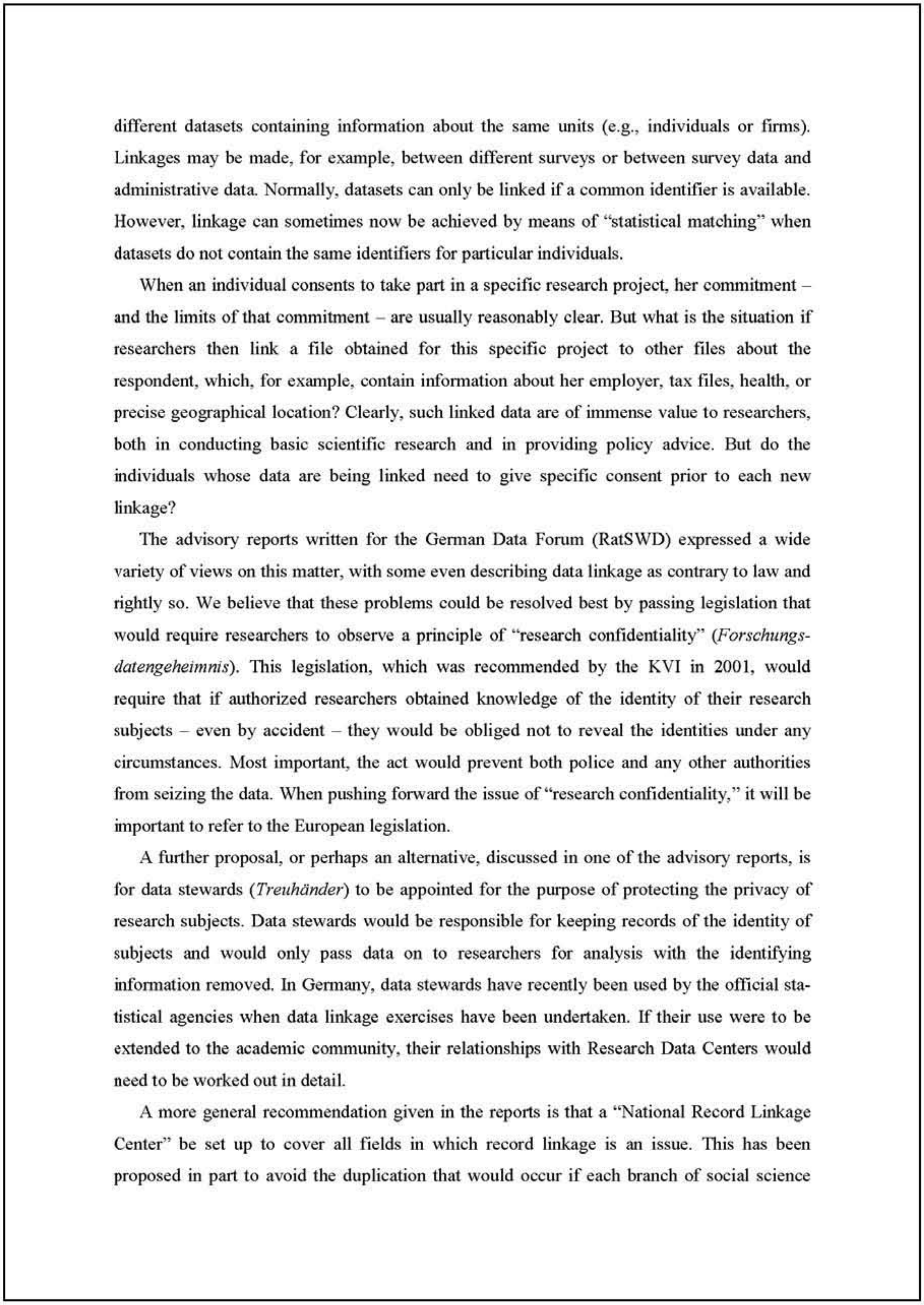
In Germany, the Research Data Centers have taken the lead in trying to improve current standards. Based on their experience, it appears that there are two internationally acceptable sets of metadata standards – the Data Documentation Initiative (DDI) and the Statistical Data and Metadata Exchange (SDMX) Standard – which could be more widely used in Germany. Adoption of these standards requires the establishment of a registry-based IT infrastructure compatible with the industry standard for Web services. This infrastructure can then facilitate the management, exchange, harmonization, and re-use of data and metadata.

We would like to highlight one potential means of improving documentation in particular: the use of a unique identifier for datasets (e.g., a digital object identifier or DOI). Unique identifiers for particular measurement scales (e.g., the different versions of the "Big Five" inventory) could possibly also be helpful (see also Theme 10 below).

The need for high-quality metadata appears even more pressing when recalling that many Internet users who are not themselves scholars are making increased use of these data for their own analyses. Results generated by lay users are especially likely to be skewed or misleading if the strengths and limitations of the data are described inadequately or in jargon a layperson could not be expected to understand.

Theme 8: Privacy issues

This section deals with privacy issues, particularly those that arise due to increasingly sophisticated methods of data linkage. *Record linkage* refers to the possibility of linking up



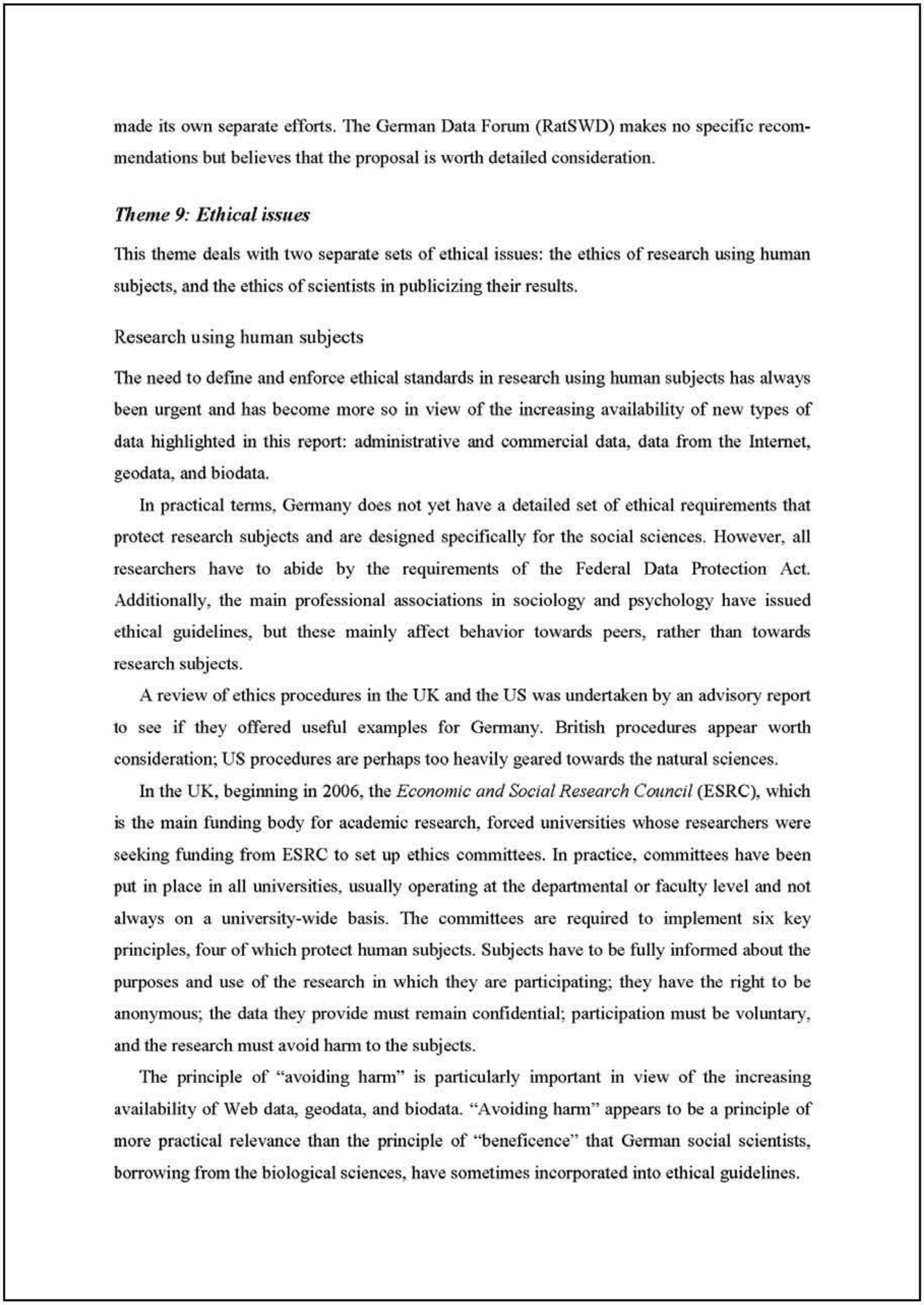
different datasets containing information about the same units (e.g., individuals or firms). Linkages may be made, for example, between different surveys or between survey data and administrative data. Normally, datasets can only be linked if a common identifier is available. However, linkage can sometimes now be achieved by means of "statistical matching" when datasets do not contain the same identifiers for particular individuals.

When an individual consents to take part in a specific research project, her commitment – and the limits of that commitment – are usually reasonably clear. But what is the situation if researchers then link a file obtained for this specific project to other files about the respondent, which, for example, contain information about her employer, tax files, health, or precise geographical location? Clearly, such linked data are of immense value to researchers, both in conducting basic scientific research and in providing policy advice. But do the individuals whose data are being linked need to give specific consent prior to each new linkage?

The advisory reports written for the German Data Forum (RatSWD) expressed a wide variety of views on this matter, with some even describing data linkage as contrary to law and rightly so. We believe that these problems could be resolved best by passing legislation that would require researchers to observe a principle of "research confidentiality" (*Forschungsdatengeheimnis*). This legislation, which was recommended by the KVI in 2001, would require that if authorized researchers obtained knowledge of the identity of their research subjects – even by accident – they would be obliged not to reveal the identities under any circumstances. Most important, the act would prevent both police and any other authorities from seizing the data. When pushing forward the issue of "research confidentiality," it will be important to refer to the European legislation.

A further proposal, or perhaps an alternative, discussed in one of the advisory reports, is for data stewards (*Treuhänder*) to be appointed for the purpose of protecting the privacy of research subjects. Data stewards would be responsible for keeping records of the identity of subjects and would only pass data on to researchers for analysis with the identifying information removed. In Germany, data stewards have recently been used by the official statistical agencies when data linkage exercises have been undertaken. If their use were to be extended to the academic community, their relationships with Research Data Centers would need to be worked out in detail.

A more general recommendation given in the reports is that a "National Record Linkage Center" be set up to cover all fields in which record linkage is an issue. This has been proposed in part to avoid the duplication that would occur if each branch of social science



made its own separate efforts. The German Data Forum (RatSWD) makes no specific recommendations but believes that the proposal is worth detailed consideration.

Theme 9: Ethical issues

This theme deals with two separate sets of ethical issues: the ethics of research using human subjects, and the ethics of scientists in publicizing their results.

Research using human subjects

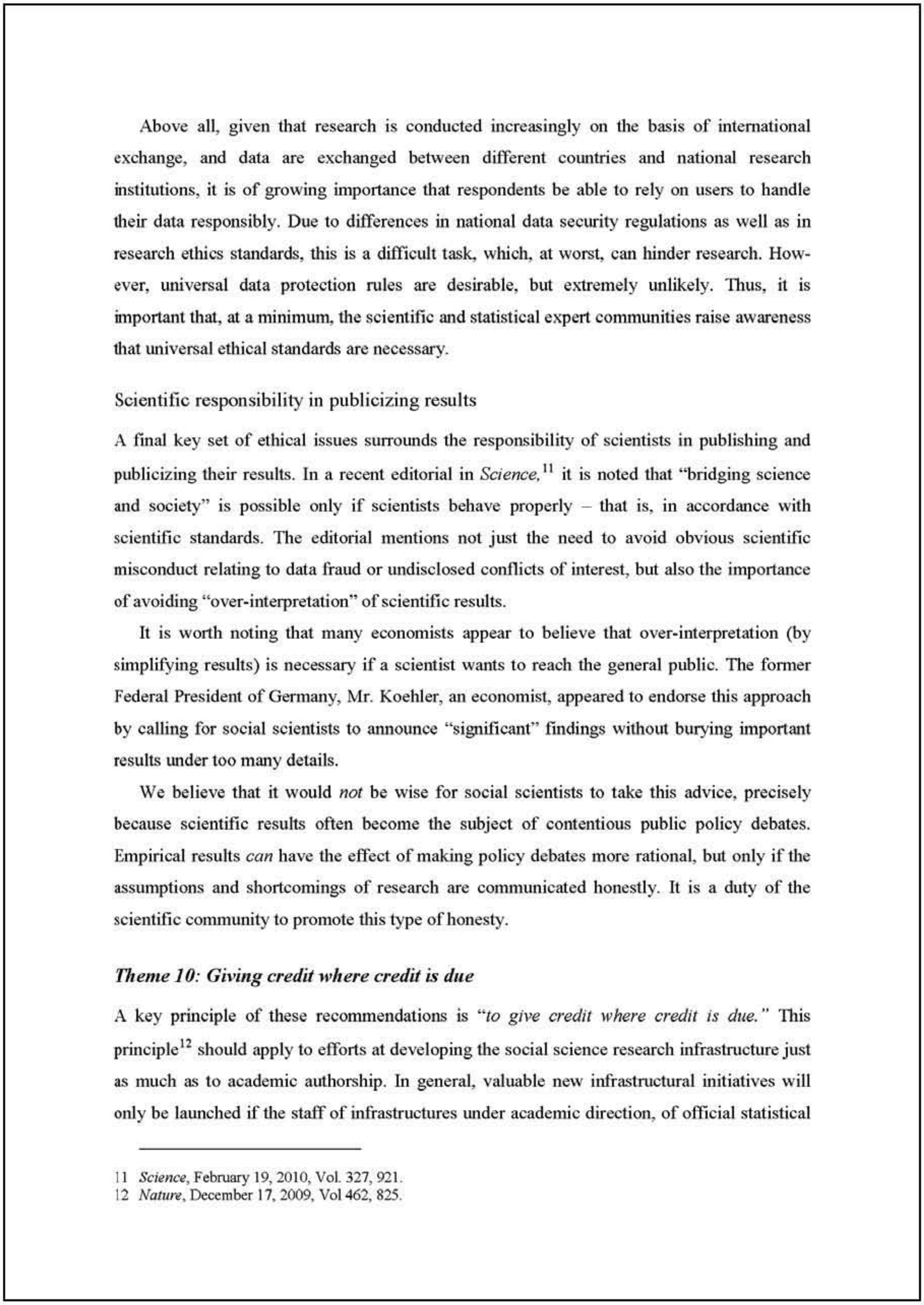
The need to define and enforce ethical standards in research using human subjects has always been urgent and has become more so in view of the increasing availability of new types of data highlighted in this report: administrative and commercial data, data from the Internet, geodata, and biodata.

In practical terms, Germany does not yet have a detailed set of ethical requirements that protect research subjects and are designed specifically for the social sciences. However, all researchers have to abide by the requirements of the Federal Data Protection Act. Additionally, the main professional associations in sociology and psychology have issued ethical guidelines, but these mainly affect behavior towards peers, rather than towards research subjects.

A review of ethics procedures in the UK and the US was undertaken by an advisory report to see if they offered useful examples for Germany. British procedures appear worth consideration; US procedures are perhaps too heavily geared towards the natural sciences.

In the UK, beginning in 2006, the *Economic and Social Research Council* (ESRC), which is the main funding body for academic research, forced universities whose researchers were seeking funding from ESRC to set up ethics committees. In practice, committees have been put in place in all universities, usually operating at the departmental or faculty level and not always on a university-wide basis. The committees are required to implement six key principles, four of which protect human subjects. Subjects have to be fully informed about the purposes and use of the research in which they are participating; they have the right to be anonymous; the data they provide must remain confidential; participation must be voluntary, and the research must avoid harm to the subjects.

The principle of "avoiding harm" is particularly important in view of the increasing availability of Web data, geodata, and biodata. "Avoiding harm" appears to be a principle of more practical relevance than the principle of "beneficence" that German social scientists, borrowing from the biological sciences, have sometimes incorporated into ethical guidelines.



Above all, given that research is conducted increasingly on the basis of international exchange, and data are exchanged between different countries and national research institutions, it is of growing importance that respondents be able to rely on users to handle their data responsibly. Due to differences in national data security regulations as well as in research ethics standards, this is a difficult task, which, at worst, can hinder research. However, universal data protection rules are desirable, but extremely unlikely. Thus, it is important that, at a minimum, the scientific and statistical expert communities raise awareness that universal ethical standards are necessary.

Scientific responsibility in publicizing results

A final key set of ethical issues surrounds the responsibility of scientists in publishing and publicizing their results. In a recent editorial in *Science,*[11] it is noted that "bridging science and society" is possible only if scientists behave properly – that is, in accordance with scientific standards. The editorial mentions not just the need to avoid obvious scientific misconduct relating to data fraud or undisclosed conflicts of interest, but also the importance of avoiding "over-interpretation" of scientific results.

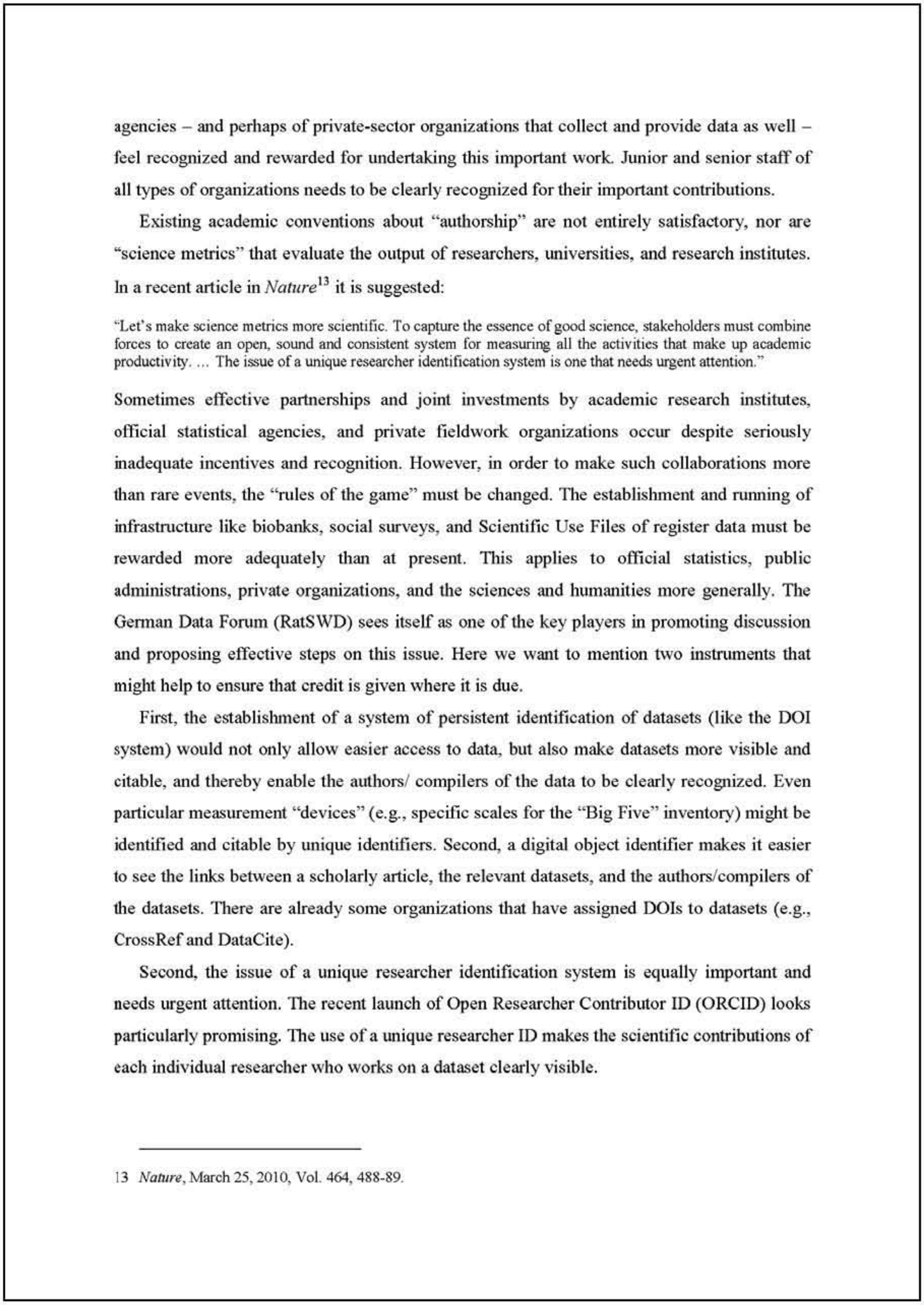
It is worth noting that many economists appear to believe that over-interpretation (by simplifying results) is necessary if a scientist wants to reach the general public. The former Federal President of Germany, Mr. Koehler, an economist, appeared to endorse this approach by calling for social scientists to announce "significant" findings without burying important results under too many details.

We believe that it would *not* be wise for social scientists to take this advice, precisely because scientific results often become the subject of contentious public policy debates. Empirical results *can* have the effect of making policy debates more rational, but only if the assumptions and shortcomings of research are communicated honestly. It is a duty of the scientific community to promote this type of honesty.

Theme 10: Giving credit where credit is due

A key principle of these recommendations is "*to give credit where credit is due.*" This principle[12] should apply to efforts at developing the social science research infrastructure just as much as to academic authorship. In general, valuable new infrastructural initiatives will only be launched if the staff of infrastructures under academic direction, of official statistical

[11] *Science*, February 19, 2010, Vol. 327, 921.
[12] *Nature*, December 17, 2009, Vol 462, 825.



agencies – and perhaps of private-sector organizations that collect and provide data as well – feel recognized and rewarded for undertaking this important work. Junior and senior staff of all types of organizations needs to be clearly recognized for their important contributions.

Existing academic conventions about "authorship" are not entirely satisfactory, nor are "science metrics" that evaluate the output of researchers, universities, and research institutes. In a recent article in *Nature*[13] it is suggested:

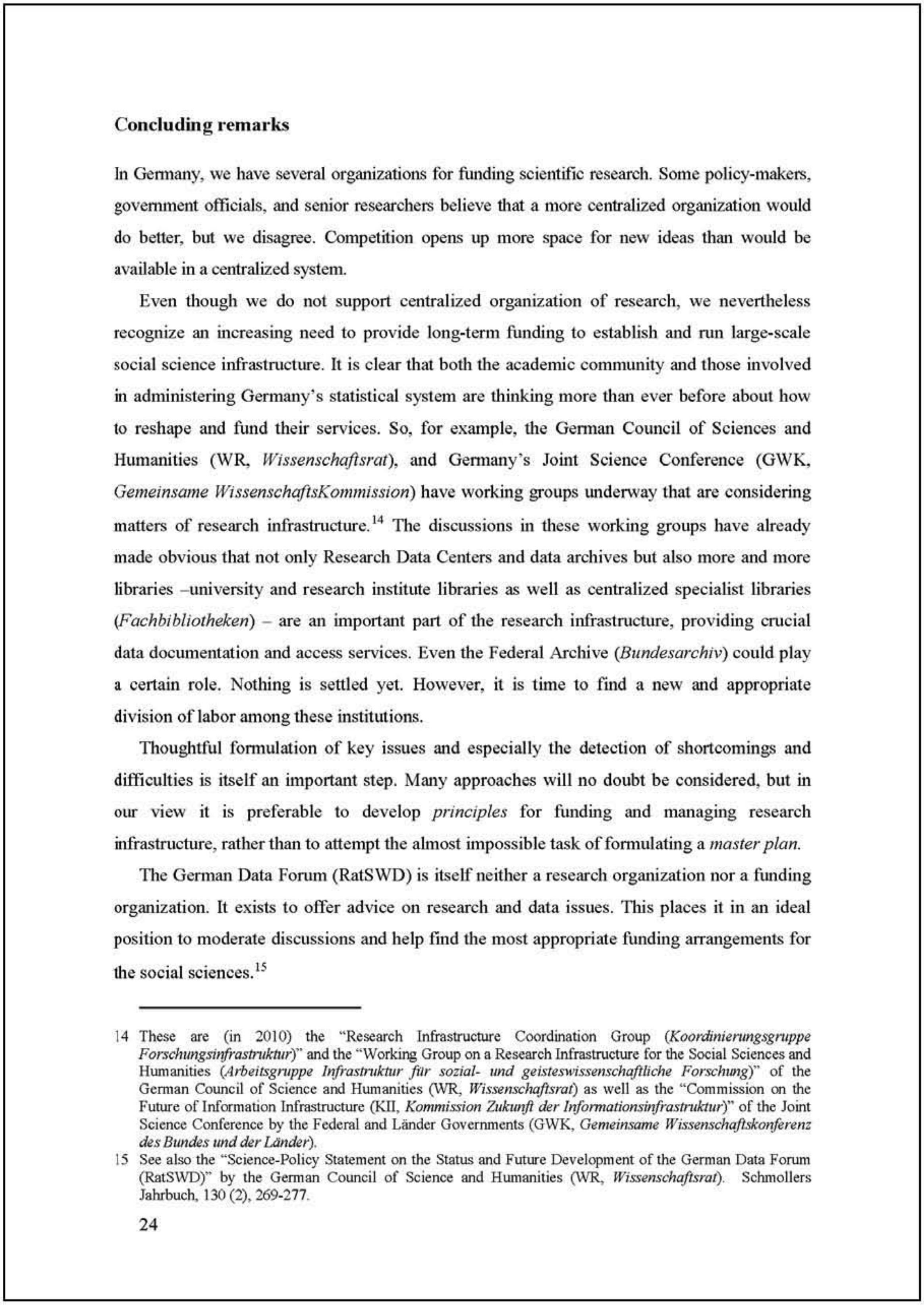
> "Let's make science metrics more scientific. To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity. ... The issue of a unique researcher identification system is one that needs urgent attention."

Sometimes effective partnerships and joint investments by academic research institutes, official statistical agencies, and private fieldwork organizations occur despite seriously inadequate incentives and recognition. However, in order to make such collaborations more than rare events, the "rules of the game" must be changed. The establishment and running of infrastructure like biobanks, social surveys, and Scientific Use Files of register data must be rewarded more adequately than at present. This applies to official statistics, public administrations, private organizations, and the sciences and humanities more generally. The German Data Forum (RatSWD) sees itself as one of the key players in promoting discussion and proposing effective steps on this issue. Here we want to mention two instruments that might help to ensure that credit is given where it is due.

First, the establishment of a system of persistent identification of datasets (like the DOI system) would not only allow easier access to data, but also make datasets more visible and citable, and thereby enable the authors/ compilers of the data to be clearly recognized. Even particular measurement "devices" (e.g., specific scales for the "Big Five" inventory) might be identified and citable by unique identifiers. Second, a digital object identifier makes it easier to see the links between a scholarly article, the relevant datasets, and the authors/compilers of the datasets. There are already some organizations that have assigned DOIs to datasets (e.g., CrossRef and DataCite).

Second, the issue of a unique researcher identification system is equally important and needs urgent attention. The recent launch of Open Researcher Contributor ID (ORCID) looks particularly promising. The use of a unique researcher ID makes the scientific contributions of each individual researcher who works on a dataset clearly visible.

[13] *Nature*, March 25, 2010, Vol. 464, 488-89.



Concluding remarks

In Germany, we have several organizations for funding scientific research. Some policy-makers, government officials, and senior researchers believe that a more centralized organization would do better, but we disagree. Competition opens up more space for new ideas than would be available in a centralized system.

Even though we do not support centralized organization of research, we nevertheless recognize an increasing need to provide long-term funding to establish and run large-scale social science infrastructure. It is clear that both the academic community and those involved in administering Germany's statistical system are thinking more than ever before about how to reshape and fund their services. So, for example, the German Council of Sciences and Humanities (WR, *Wissenschaftsrat*), and Germany's Joint Science Conference (GWK, *Gemeinsame WissenschaftsKommission*) have working groups underway that are considering matters of research infrastructure.[14] The discussions in these working groups have already made obvious that not only Research Data Centers and data archives but also more and more libraries –university and research institute libraries as well as centralized specialist libraries (*Fachbibliotheken*) – are an important part of the research infrastructure, providing crucial data documentation and access services. Even the Federal Archive (*Bundesarchiv*) could play a certain role. Nothing is settled yet. However, it is time to find a new and appropriate division of labor among these institutions.

Thoughtful formulation of key issues and especially the detection of shortcomings and difficulties is itself an important step. Many approaches will no doubt be considered, but in our view it is preferable to develop *principles* for funding and managing research infrastructure, rather than to attempt the almost impossible task of formulating a *master plan.*

The German Data Forum (RatSWD) is itself neither a research organization nor a funding organization. It exists to offer advice on research and data issues. This places it in an ideal position to moderate discussions and help find the most appropriate funding arrangements for the social sciences.[15]

[14] These are (in 2010) the "Research Infrastructure Coordination Group (*Koordinierungsgruppe Forschungsinfrastruktur*)" and the "Working Group on a Research Infrastructure for the Social Sciences and Humanities (*Arbeitsgruppe Infrastruktur für sozial- und geisteswissenschaftliche Forschung*)" of the German Council of Science and Humanities (WR, *Wissenschaftsrat*) as well as the "Commission on the Future of Information Infrastructure (KII, *Kommission Zukunft der Informationsinfrastruktur*)" of the Joint Science Conference by the Federal and Länder Governments (GWK, *Gemeinsame Wissenschaftskonferenz des Bundes und der Länder*).

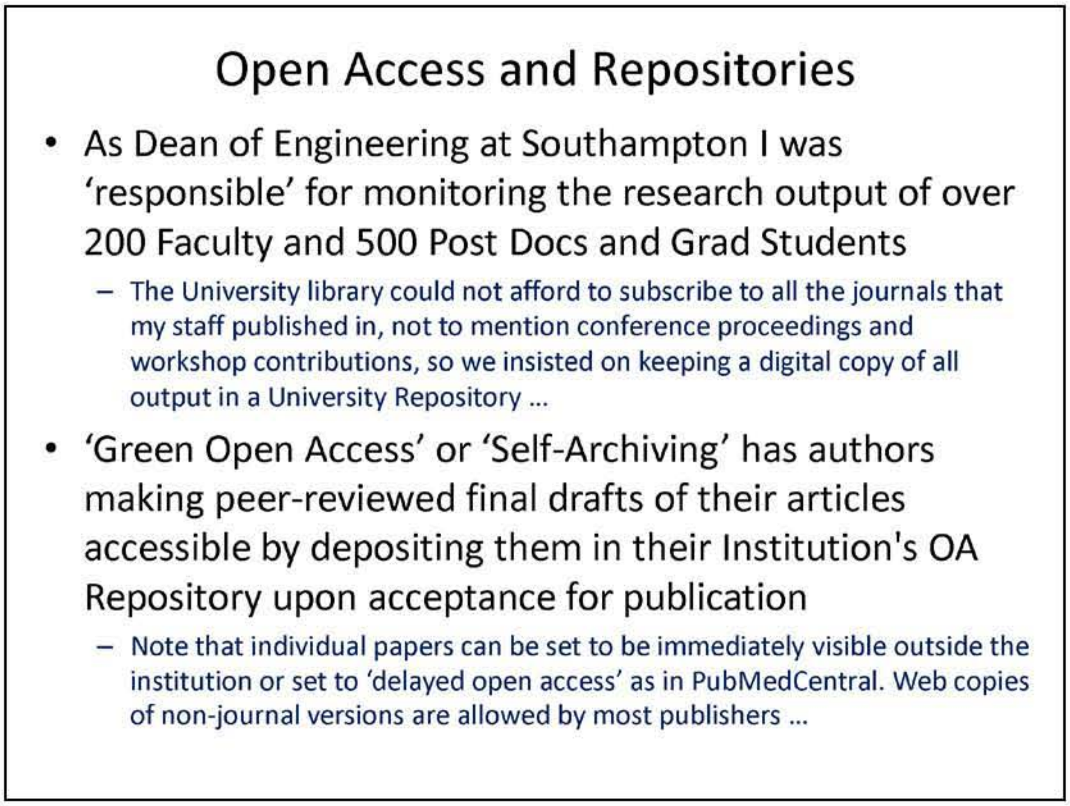
[15] See also the "Science-Policy Statement on the Status and Future Development of the German Data Forum (RatSWD)" by the German Council of Science and Humanities (WR, *Wissenschaftsrat*). Schmollers Jahrbuch, 130 (2), 269-277.

Hey, Tony, “Open Access, Open Data, Open Science”

Open Access, Open Data, Open Science

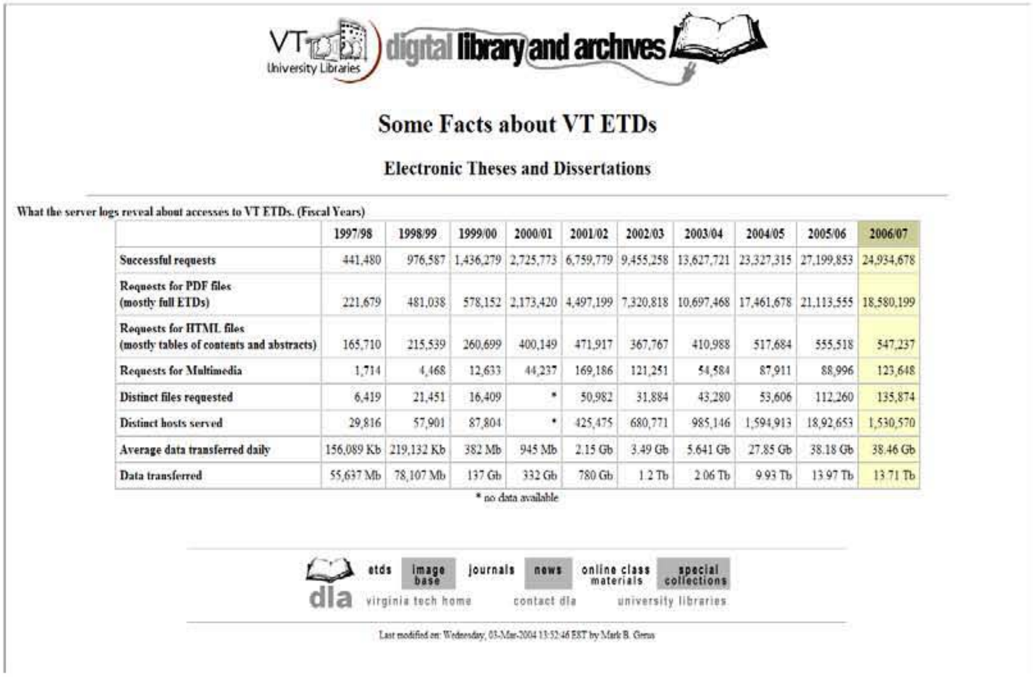
Tony Hey

Microsoft Research



Open Access and Repositories

- As Dean of Engineering at Southampton I was ‘responsible’ for monitoring the research output of over 200 Faculty and 500 Post Docs and Grad Students
  - The University library could not afford to subscribe to all the journals that my staff published in, not to mention conference proceedings and workshop contributions, so we insisted on keeping a digital copy of all output in a University Repository …
- ‘Green Open Access’ or ‘Self-Archiving’ has authors making peer-reviewed final drafts of their articles accessible by depositing them in their Institution's OA Repository upon acceptance for publication
  - Note that individual papers can be set to be immediately visible outside the institution or set to ‘delayed open access’ as in PubMedCentral. Web copies of non-journal versions are allowed by most publishers …



Some Facts about VT ETDs

Electronic Theses and Dissertations

What the server logs reveal about accesses to VT ETDs. (Fiscal Years)

	1997/98	1998/99	1999/00	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07
Successful requests	441,480	976,587	1,436,279	2,725,773	6,759,779	9,455,258	13,627,721	23,327,315	27,199,853	24,934,678
Requests for PDF files (mostly full ETDs)	221,679	481,038	578,152	2,173,420	4,497,199	7,320,818	10,697,468	17,461,678	21,113,555	18,580,199
Requests for HTML files (mostly tables of contents and abstracts)	165,710	215,539	260,699	400,149	471,917	367,767	410,988	517,684	555,518	547,237
Requests for Multimedia	1,714	4,468	12,633	44,237	169,186	121,251	54,584	87,911	88,996	123,648
Distinct files requested	6,419	21,451	16,409	*	50,982	31,884	43,280	53,606	112,260	135,874
Distinct hosts served	29,816	57,901	87,804	*	425,475	680,771	985,146	1,594,913	18,92,653	1,530,570
Average data transferred daily	156,089 Kb	219,132 Kb	382 Mb	945 Mb	2.15 Gb	3.49 Gb	5.641 Gb	27.85 Gb	38.18 Gb	38.46 Gb
Data transferred	55,637 Mb	78,107 Mb	137 Gb	332 Gb	780 Gb	1.2 Tb	2.06 Tb	9.93 Tb	13.97 Tb	13.71 Tb

* no data available

Last modified on: Wednesday, 03-Mar-2004 13:52:46 EST by Mark B. Gerns

- ➢ Demonstrates the Power of the Web

Webometrics Google Scholar Ranking (July 2010)

1	Harvard
2	MIT
3	UNAM
4	Minnesota
5	UC Madrid
6	Munich
7	Stanford
8	U Queensland
9	Kyoto
10	Masaryk
11	Toronto
12	Michigan
13	UPC Barcelona
14	Texas A&M
15	ETH Zurich
16	Nebraska
17	Groningen
18	Vienna
19	CUHK
20	Georgia Tech
21	Southampton
22	Cornell
23	Pennsylvania
24	Tokyo
25	Murcia

Southampton	# 21
VirginiaTech	# 37
Cambridge	# 97
Oxford	# 115

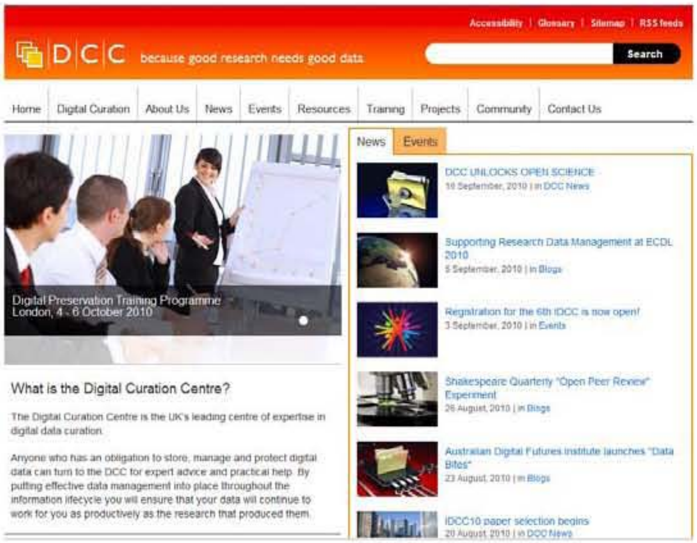
Clearly not a 'perfect' metric - but equally clearly, this must measure something of relevance for the research reputation of a university …

- ➢ Institutional Research Repository must be part of the university's '**Reputation Management**' strategy

Six Key Elements for a Global Cyberinfrastructure for eScience (2004)

1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centers for Open Software
4. **Technologies and standards for Data Provenance, Curation and Preservation**
5. **Open access to Data and Publications via Interoperable Repositories**
6. Discovery Services and Collaborative Tools

UK Digital Curation Centre (JISC funded 2004)


Accessibility | Glossary | Sitemap | RSS feeds
DCC because good research needs good data
Search
Home
Digital Curation
About Us
News
Events
Resources
Training
Projects
Community
Contact Us
Digital Preservation Training Programme
London, 4 - 6 October 2010
News
Events

What is the Digital Curation Centre?

The Digital Curation Centre is the UK's leading centre of expertise in digital data curation.

Anyone who has an obligation to store, manage and protect digital data can turn to the DCC for expert advice and practical help. By putting effective data management into place throughout the information lifecycle you will ensure that your data will continue to work for you as productively as the research that produced them.

DCC UNLOCKS OPEN SCIENCE
10 September, 2010 | in DCC News

Supporting Research Data Management at ECDL 2010
5 September, 2010 | in Blogs

Registration for the 6th IDCC is now open!
3 September, 2010 | in Events

Shakespeare Quarterly "Open Peer Review" Experiment
26 August, 2010 | in Blogs

Australian Digital Futures Institute launches "Data Bites"
23 August, 2010 | in Blogs

IDCC10 paper selection begins
20 August, 2010 | in DCC News

http://www.dcc.ac.uk

Jim Gray's Call to Action

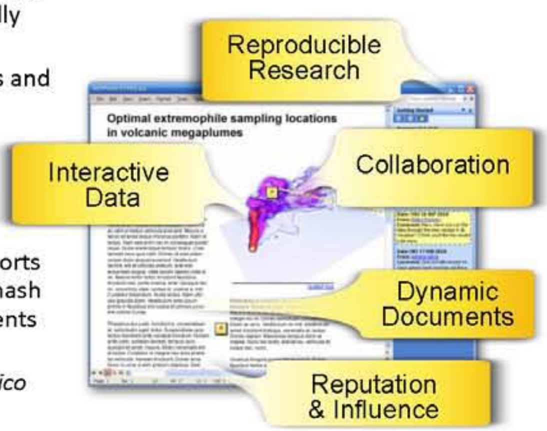
In his last talk Jim Gray highlighted three key areas for action relating to the future of Scholarly Communication and Libraries:

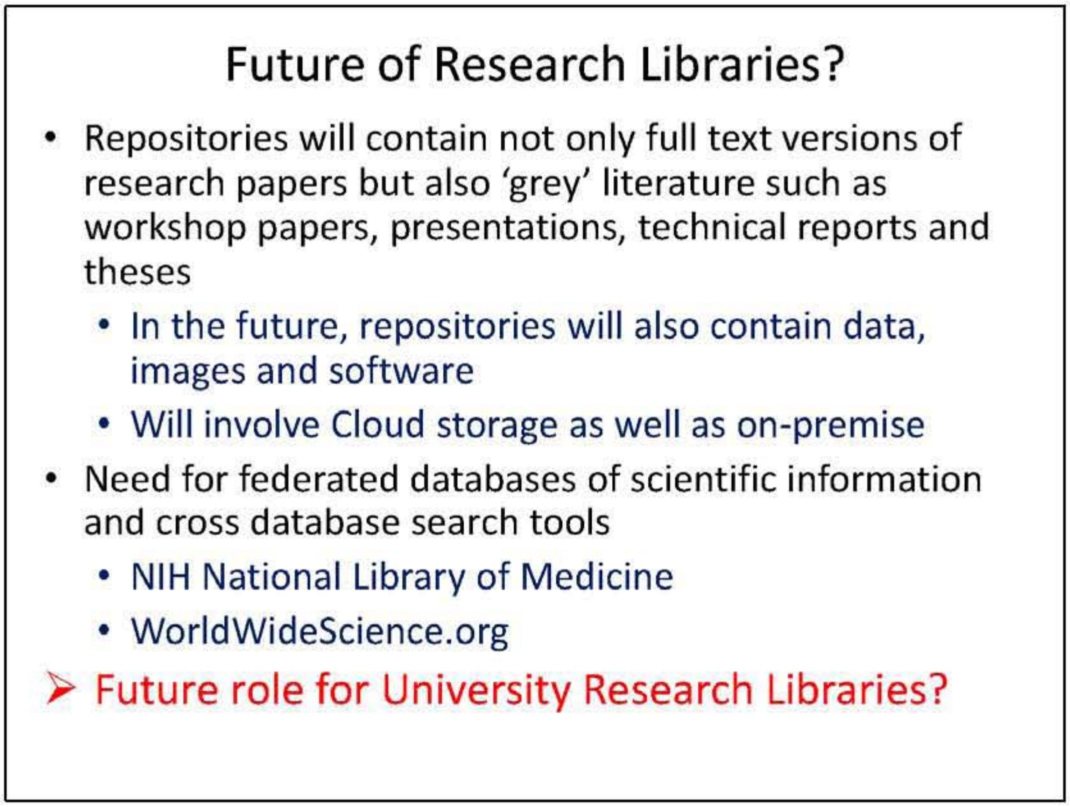
1. Establish Digital Libraries that support the other sciences like the NLM does for Medicine
2. Fund development of new authoring tools and publication models
3. Explore development of digital data libraries that contain scientific data (not just the metadata) and support integration with published literature

Envisioning a New Era of Research Reporting

Imagine…

- Live research reports that had multiple end-user 'views' and which could dynamically tailor their presentation to each user
- An authoring environment that absorbs and encapsulates research workflows and outputs from the lab experiments
- A report that can be dropped into an electronic lab workbench in order to reconstitute an entire experiment
- A researcher working with multiple reports on a Surface and having the ability to mash up data and workflows across experiments
- The ability to apply new analyses and visualizations and to perform new *in silico* experiments

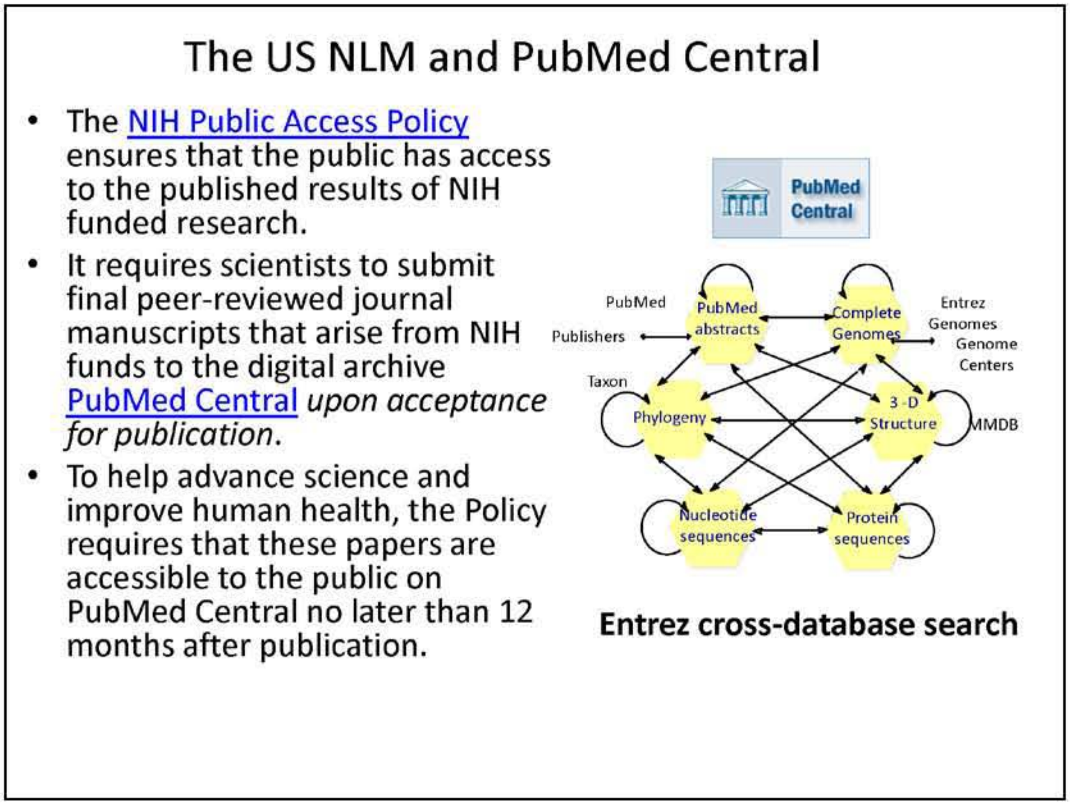




Future of Research Libraries?

- Repositories will contain not only full text versions of research papers but also 'grey' literature such as workshop papers, presentations, technical reports and theses
  - In the future, repositories will also contain data, images and software
  - Will involve Cloud storage as well as on-premise
- Need for federated databases of scientific information and cross database search tools
  - NIH National Library of Medicine
  - WorldWideScience.org

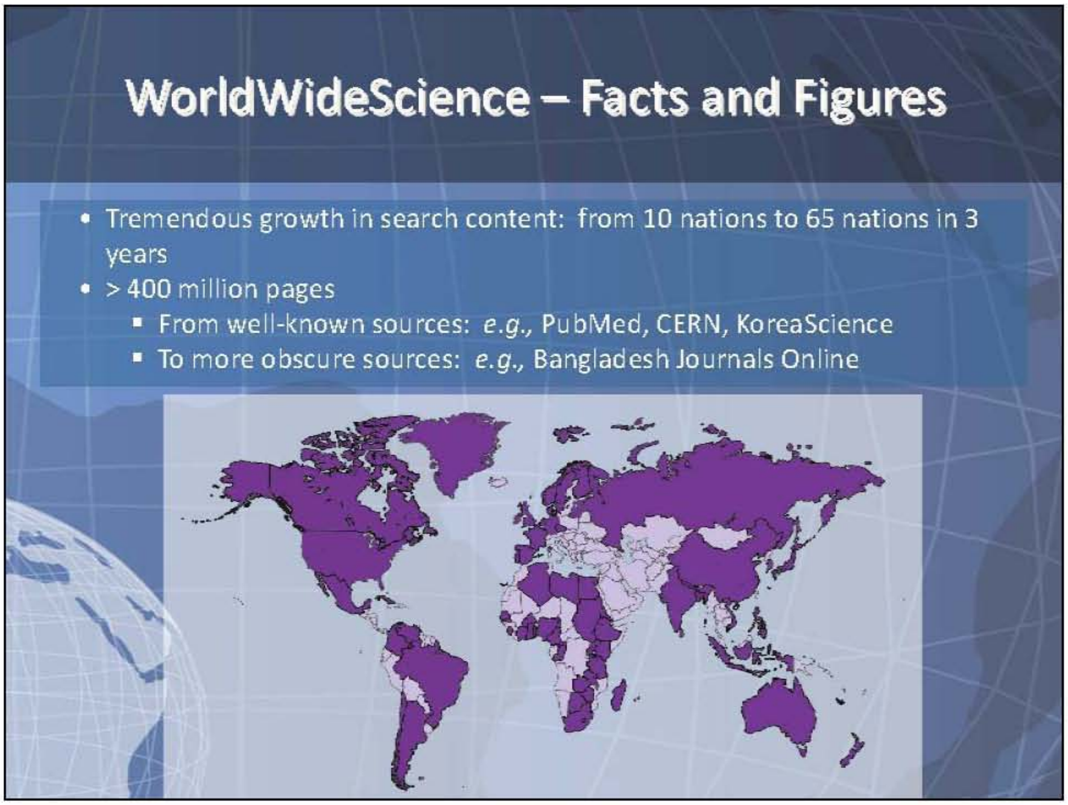
➢ Future role for University Research Libraries?



The US NLM and PubMed Central

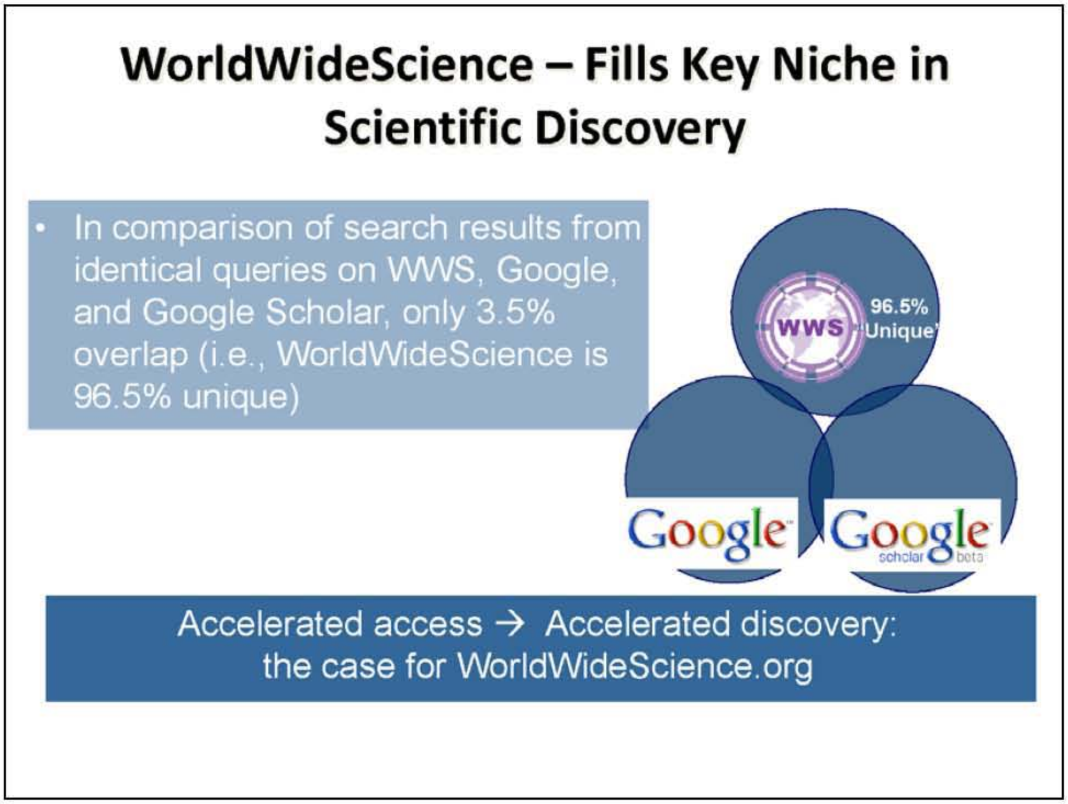
- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

Entrez cross-database search



WorldWideScience – Facts and Figures

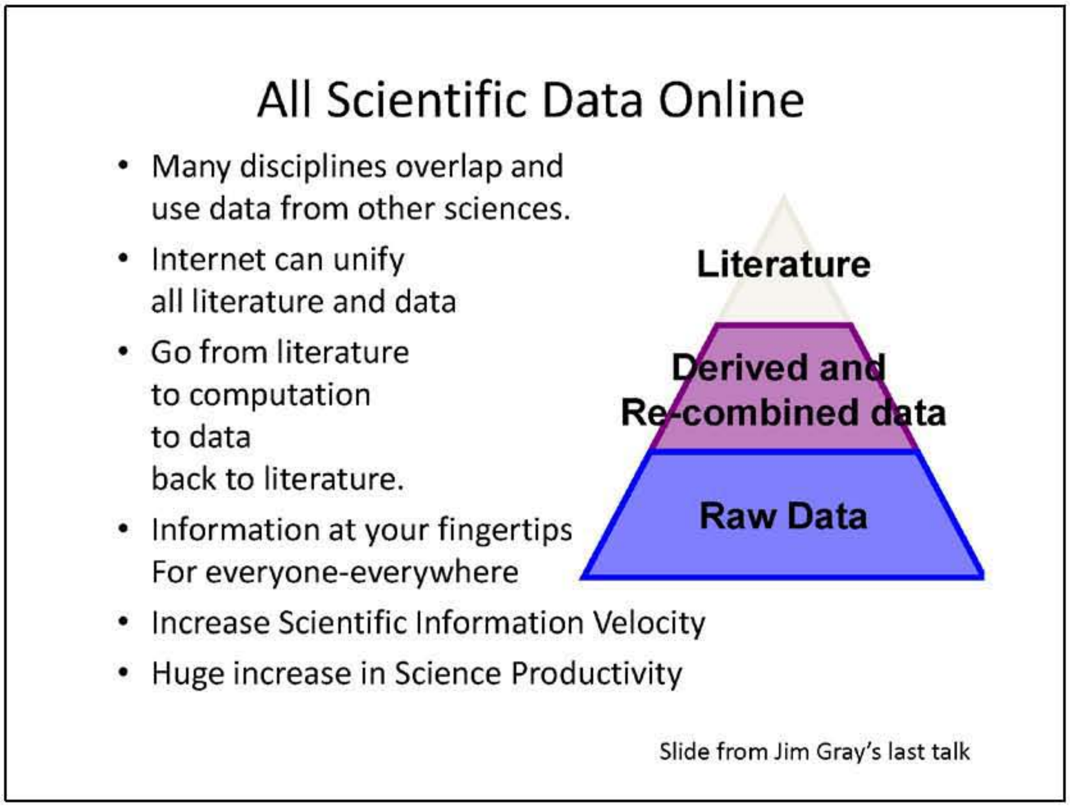
- Tremendous growth in search content: from 10 nations to 65 nations in 3 years
- > 400 million pages
  - From well-known sources: *e.g.*, PubMed, CERN, KoreaScience
  - To more obscure sources: *e.g.*, Bangladesh Journals Online



WorldWideScience – Fills Key Niche in Scientific Discovery

- In comparison of search results from identical queries on WWS, Google, and Google Scholar, only 3.5% overlap (i.e., WorldWideScience is 96.5% unique)

Accelerated access → Accelerated discovery: the case for WorldWideScience.org



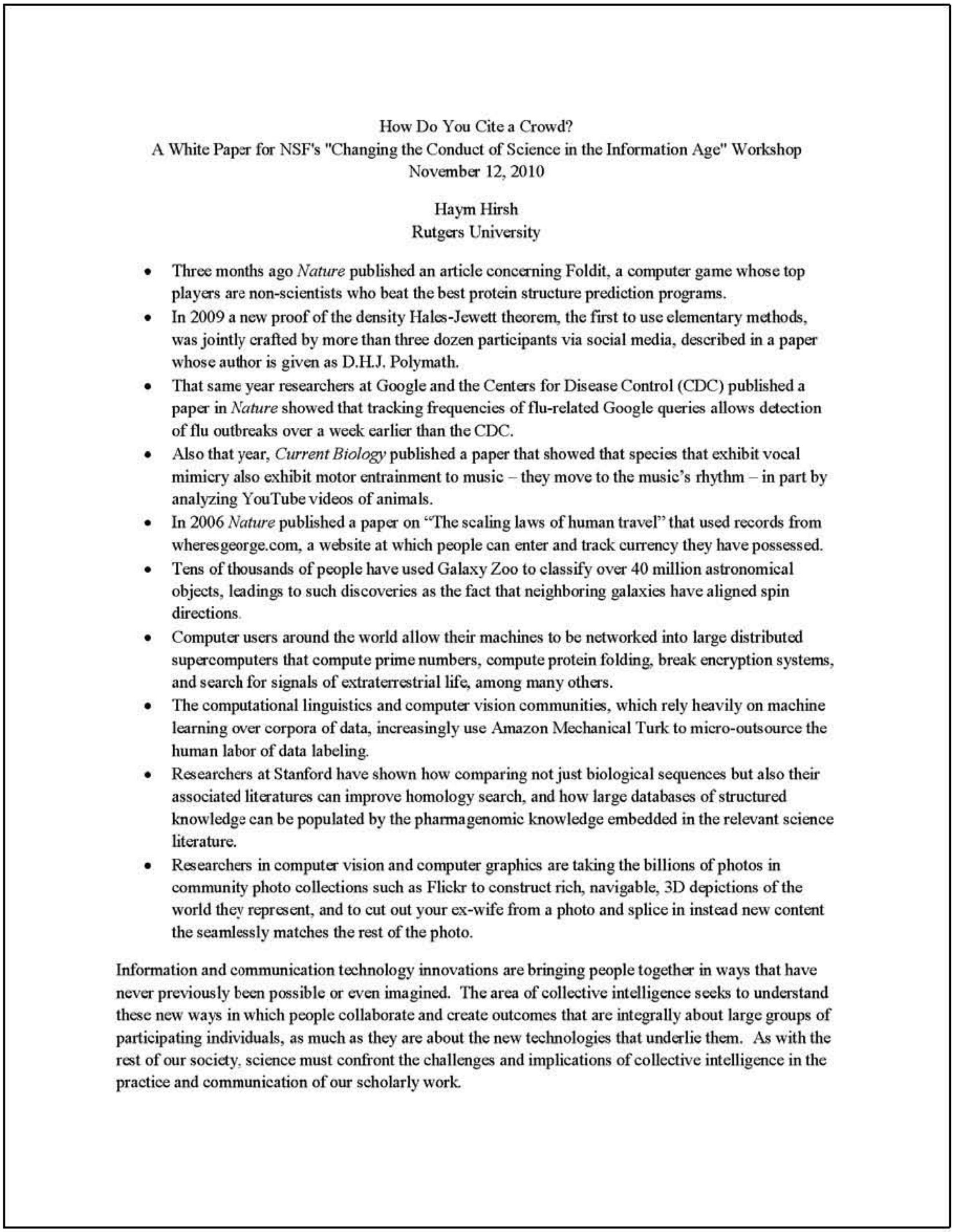
All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature
  to computation
  to data
  back to literature.
- Information at your fingertips
  For everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity

Literature
Derived and Re-combined data
Raw Data

Slide from Jim Gray's last talk

Hirsh, Haym, “How Do You Cite a Crowd?”

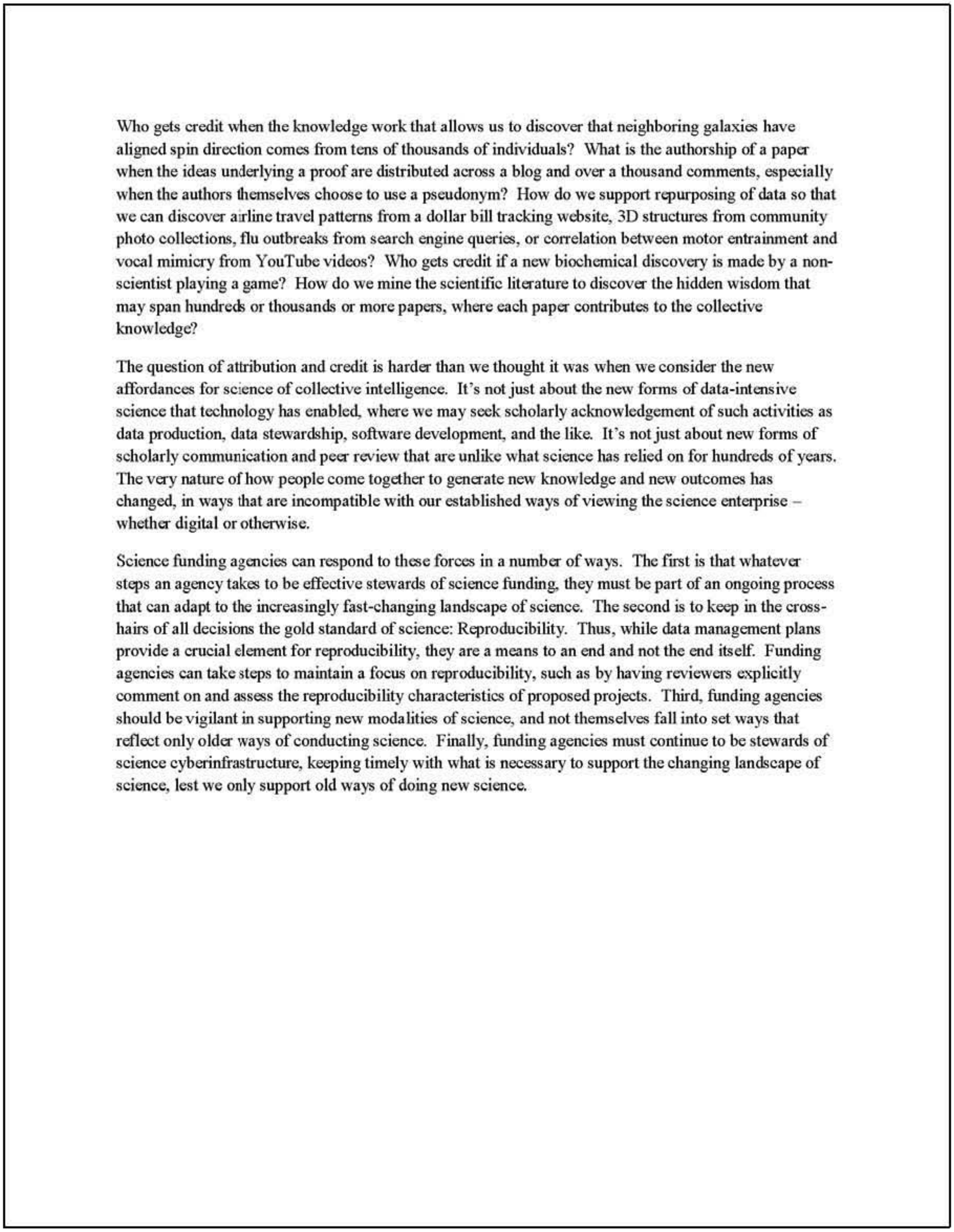


How Do You Cite a Crowd?
A White Paper for NSF's "Changing the Conduct of Science in the Information Age" Workshop
November 12, 2010

Haym Hirsh
Rutgers University

- Three months ago *Nature* published an article concerning Foldit, a computer game whose top players are non-scientists who beat the best protein structure prediction programs.
- In 2009 a new proof of the density Hales-Jewett theorem, the first to use elementary methods, was jointly crafted by more than three dozen participants via social media, described in a paper whose author is given as D.H.J. Polymath.
- That same year researchers at Google and the Centers for Disease Control (CDC) published a paper in *Nature* showed that tracking frequencies of flu-related Google queries allows detection of flu outbreaks over a week earlier than the CDC.
- Also that year, *Current Biology* published a paper that showed that species that exhibit vocal mimicry also exhibit motor entrainment to music – they move to the music's rhythm – in part by analyzing YouTube videos of animals.
- In 2006 *Nature* published a paper on “The scaling laws of human travel” that used records from wheresgeorge.com, a website at which people can enter and track currency they have possessed.
- Tens of thousands of people have used Galaxy Zoo to classify over 40 million astronomical objects, leadings to such discoveries as the fact that neighboring galaxies have aligned spin directions.
- Computer users around the world allow their machines to be networked into large distributed supercomputers that compute prime numbers, compute protein folding, break encryption systems, and search for signals of extraterrestrial life, among many others.
- The computational linguistics and computer vision communities, which rely heavily on machine learning over corpora of data, increasingly use Amazon Mechanical Turk to micro-outsource the human labor of data labeling.
- Researchers at Stanford have shown how comparing not just biological sequences but also their associated literatures can improve homology search, and how large databases of structured knowledge can be populated by the pharmagenomic knowledge embedded in the relevant science literature.
- Researchers in computer vision and computer graphics are taking the billions of photos in community photo collections such as Flickr to construct rich, navigable, 3D depictions of the world they represent, and to cut out your ex-wife from a photo and splice in instead new content the seamlessly matches the rest of the photo.

Information and communication technology innovations are bringing people together in ways that have never previously been possible or even imagined. The area of collective intelligence seeks to understand these new ways in which people collaborate and create outcomes that are integrally about large groups of participating individuals, as much as they are about the new technologies that underlie them. As with the rest of our society, science must confront the challenges and implications of collective intelligence in the practice and communication of our scholarly work.

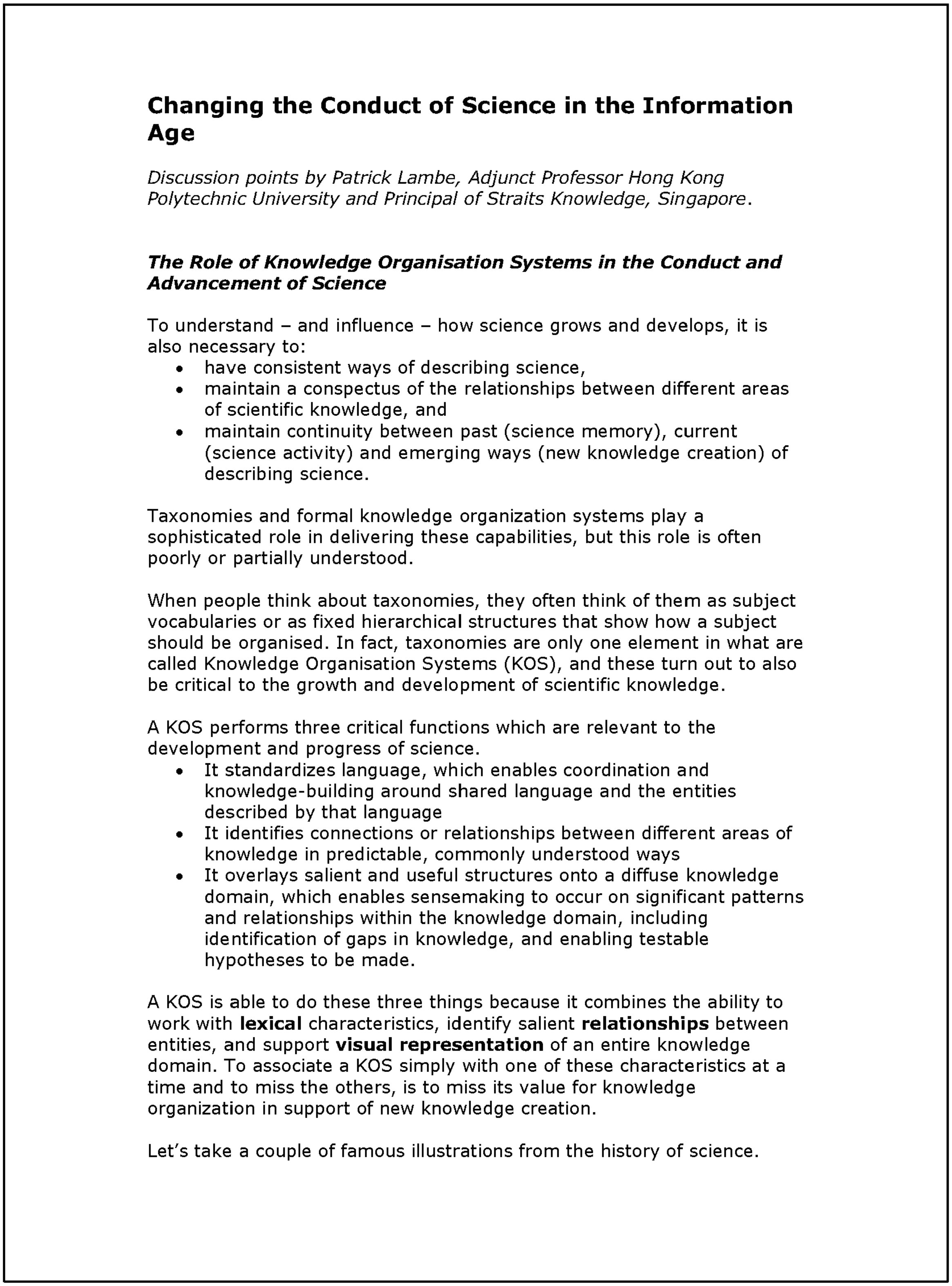


Who gets credit when the knowledge work that allows us to discover that neighboring galaxies have aligned spin direction comes from tens of thousands of individuals? What is the authorship of a paper when the ideas underlying a proof are distributed across a blog and over a thousand comments, especially when the authors themselves choose to use a pseudonym? How do we support repurposing of data so that we can discover airline travel patterns from a dollar bill tracking website, 3D structures from community photo collections, flu outbreaks from search engine queries, or correlation between motor entrainment and vocal mimicry from YouTube videos? Who gets credit if a new biochemical discovery is made by a non-scientist playing a game? How do we mine the scientific literature to discover the hidden wisdom that may span hundreds or thousands or more papers, where each paper contributes to the collective knowledge?

The question of attribution and credit is harder than we thought it was when we consider the new affordances for science of collective intelligence. It's not just about the new forms of data-intensive science that technology has enabled, where we may seek scholarly acknowledgement of such activities as data production, data stewardship, software development, and the like. It's not just about new forms of scholarly communication and peer review that are unlike what science has relied on for hundreds of years. The very nature of how people come together to generate new knowledge and new outcomes has changed, in ways that are incompatible with our established ways of viewing the science enterprise – whether digital or otherwise.

Science funding agencies can respond to these forces in a number of ways. The first is that whatever steps an agency takes to be effective stewards of science funding, they must be part of an ongoing process that can adapt to the increasingly fast-changing landscape of science. The second is to keep in the cross-hairs of all decisions the gold standard of science: Reproducibility. Thus, while data management plans provide a crucial element for reproducibility, they are a means to an end and not the end itself. Funding agencies can take steps to maintain a focus on reproducibility, such as by having reviewers explicitly comment on and assess the reproducibility characteristics of proposed projects. Third, funding agencies should be vigilant in supporting new modalities of science, and not themselves fall into set ways that reflect only older ways of conducting science. Finally, funding agencies must continue to be stewards of science cyberinfrastructure, keeping timely with what is necessary to support the changing landscape of science, lest we only support old ways of doing new science.

Lambe, Patrick, "Changing the Conduct of Science in the Information Age: Discussion Points"



Changing the Conduct of Science in the Information Age

Discussion points by Patrick Lambe, Adjunct Professor Hong Kong Polytechnic University and Principal of Straits Knowledge, Singapore.

The Role of Knowledge Organisation Systems in the Conduct and Advancement of Science

To understand – and influence – how science grows and develops, it is also necessary to:

- have consistent ways of describing science,
- maintain a conspectus of the relationships between different areas of scientific knowledge, and
- maintain continuity between past (science memory), current (science activity) and emerging ways (new knowledge creation) of describing science.

Taxonomies and formal knowledge organization systems play a sophisticated role in delivering these capabilities, but this role is often poorly or partially understood.

When people think about taxonomies, they often think of them as subject vocabularies or as fixed hierarchical structures that show how a subject should be organised. In fact, taxonomies are only one element in what are called Knowledge Organisation Systems (KOS), and these turn out to also be critical to the growth and development of scientific knowledge.

A KOS performs three critical functions which are relevant to the development and progress of science.

- It standardizes language, which enables coordination and knowledge-building around shared language and the entities described by that language
- It identifies connections or relationships between different areas of knowledge in predictable, commonly understood ways
- It overlays salient and useful structures onto a diffuse knowledge domain, which enables sensemaking to occur on significant patterns and relationships within the knowledge domain, including identification of gaps in knowledge, and enabling testable hypotheses to be made.

A KOS is able to do these three things because it combines the ability to work with **lexical** characteristics, identify salient **relationships** between entities, and support **visual representation** of an entire knowledge domain. To associate a KOS simply with one of these characteristics at a time and to miss the others, is to miss its value for knowledge organization in support of new knowledge creation.

Let's take a couple of famous illustrations from the history of science.

Carl Linnaeus

Throughout the fifteenth century, with the spreading of wealth through trade and the growth of scholarship, the passion for collecting "curiosities" was taken up on a large scale by scholars and scientists across Europe, and their collections were increasingly used as instruments of learning about the natural world. Arrangements of curiosities became part of a larger endeavour to construct a systematic knowledge of the natural world. Collections started to become more systematic and supportive of enquiry, sensemaking and discovery.

These were the seeds of modern empirical science. By the beginning of the seventeenth century, however, writers like Francis Bacon were thoroughly dismissive of the higgledy-piggledy arrangements of the rich and famous:

"There is such a multitude and host as it were of particular objects, and lying so widely dispersed, as to distract and confuse the understanding; and we can therefore hope for no advantage … unless we put its forces in due order and array by means of proper, and well arranged, and as it were living tables of discovery of these matters which are the subject of investigation…"

Bacon's impatience was echoed just over a century later by the methodical biologist Carl Linnaeus who was dismissive of the "complete disorder" he found in the home of the last great universal collector of his time, Sir Hans Sloane – founder of the collection that became the British Museum. After Sloane, in fact, collectors divided themselves into discrete disciplines. The world of knowledge had become too complex to comprehend and represent in one single arrangement.

In the midst of this complexity, Linnaeus' great gift to science was threefold. Beginning with his *Systema Natura* in 1735, he introduced a far simpler principle of distinguishing between species based on anatomical observation than had ever been proposed before. Beginning in 1737 with his *Critica Botanica* he laid down the rules for his binomial naming system for species which riled his critics immensely (because he substituted so many older naming conventions with his own), but when widely adopted created the first standardized way of describing species. This immeasurably enhanced scientific coordination and collaboration.

Finally, his hierarchical, nested classification tree structure turned out to be a perfect vehicle to express the genealogical relationships that gained such prominence during the emerging evolutionary theories of the late eighteenth and early nineteenth centuries.

Linnaeus' new taxonomic method simplified the task of categorization, imposed rigorous rules (and therefore consistency), and happened on a form of representation that history turned into a lucky bet. From the point of view of advancing scientific method, his focus on analysis, rules and standardized approaches, gave an incalculable advantage.

We can see in Linnaeus' taxonomy design two of the three elements of a

KOS – lexical stabilization to enable coordination between scientists, and a meaningful structure (a hierarchical rule-based tree structure) to establish predictable and (as it turned out from subsequent science) salient relationships between the entities being described.

Dmitri Mendeleev

Dmitri Mendeleev's periodic table of elements was an attempt to figure out patterns of behaviour across chemical elements. His endeavour was essentially a sensemaking endeavour illustrating the third function of a KOS – he was playing with the organization of the elements to see if he could explain deviations, simplify, understand and explain the relationships between them.

Mendeleev used a different taxonomy structure, not the classical hierarchy associated with Linnaeus. He used the matrix structure, where the entities are arranged according to their properties along two dimensions –he arranged the elements in columns by similarity of properties and horizontally by regular patterns of behaviour or periodicity. Like Linnaeus, he happened upon a salient and useful way of organizing before the underlying science behind his arrangement had been uncovered – electron structures had not yet been identified.

Arranging the elements in this way did two interesting things for science. First, it helped to make sense of the "periodicity" of elements – where elements exhibit similar properties at regular intervals of atomic mass increase. Secondly, representing the elements in a matrix display enabled scientists to identify gaps in the table where elements that were previously unknown should exist.

Hence the KOS helped explain behaviours and gave predictive power by identifying new elements that scientists could hunt for – and were subsequently discovered or manufactured in the laboratory – simply because their "place" in the taxonomy was visibly unfilled. Discovering and displaying the periodicity of behaviour through organizing by mass and electron structure allowed scientists to predict the existence of new elements – essentially to create new knowledge.

This by the way turns out to be a strong feature of matrix representations for taxonomies. They are extremely useful for sensemaking as well as for new knowledge creation or discovery.

Linnaeus and Mendeleev created knowledge organisation systems and standardised scientific languages to enable greater coordination, inter-connection and sensemaking across their respective scientific communities.

The elements of a KOS

A KOS can have three different orders of complexity. As science becomes more complex and inter-related, the complexity of the needed KOS increases:

(a) At the most basic level are **controlled vocabularies**, with principles for recognition, inclusion and exclusion, which provide a common reference language for describing science and enabling coordination.

(b) Next in order of complexity are **taxonomies** which put structure around the controlled vocabularies (along with principles for how those structures are maintained), and which enable sensemaking, identification of gaps, and inter-relationships among areas of science.

(c) As scientific knowledge becomes even more complex, taxonomies can no longer represent all of the salient kinds of relationships within a single comprehensible structure. We need ways of visualizing different patterns of relationships across multiple domains. **Ontologies** are systems of taxonomies, where relationships are also defined across different taxonomies, taxonomy elements and vocabularies. They enable large scale pattern-sensing and sophisticated interpretation filters on a complex scientific activity landscape.

(d) Finally, a knowledge organisation system requires mechanisms for detecting and recognising new language, new usages and new relationships between areas of science. This is essential to keeping the KOS vocabularies, taxonomies and ontologies current and reflective of current and emerging reality. The maturing field of **topic maps** based on semantic analysis is an important example of such a mechanism.

Principle 1: the complexity of a KOS needs to match the complexity of the domain it attempts to describe, and the complexity of the coordination, connection and sensemaking work it needs to support.

Human factors in using KOS

Modern science is now too fluid and complex to be supported by simpler KOS's such as controlled vocabularies and taxonomies. This is why keyword or topic-based approaches, or single taxonomy approaches to the description and measurement of science have inherent limitations by themselves. Any controlled vocabularies in use, and any taxonomy systems in use, really need the richer environment of ontologies behind them, to perform the sensemaking, memory and coordination functions that a KOS should properly provide for the complex and shifting landscape of science.

One of the drawbacks with ontologies however is that machines find it much easier to navigate and process the information from ontologies than humans do. Humans have significant cognitive constraints in terms of attention, memory span and tracking relationships, which means that they are much more suited to navigating and processing individual taxonomies than multi-dimensional ontologies.

This has implications for the human users of a KOS who tend to favor simpler lexical work (eg keywords or topic words) or simplistic taxonomy structures over investment in the information enrichment required to support ontologies. Actors such as publishers, authors, audiences, scientists, science administrators, funders, analysts, policy makers, all require human-scale representations of scientific knowledge – and this means at the vocabulary level, or at the taxonomy level.

If ontologies are to support the human actors in the science landscape, ontologies require context-sensitive human interfaces to create intelligible representations that are meaningful to their respective audiences, but still provide those functions of standardization of language, meaningful connections of content (including from past to future), and sensemaking capability. Vocabularies need to be connected to taxonomies, and taxonomies need to be connected to ontologies.

Principle 2: when the complexity of the KOS exceeds human cognitive capabilities, designed interfaces using taxonomies are necessary to serve the working needs of users in their own normal working contexts.

Humans also resist lexical control, especially if the controlled language is not natural to their own context.

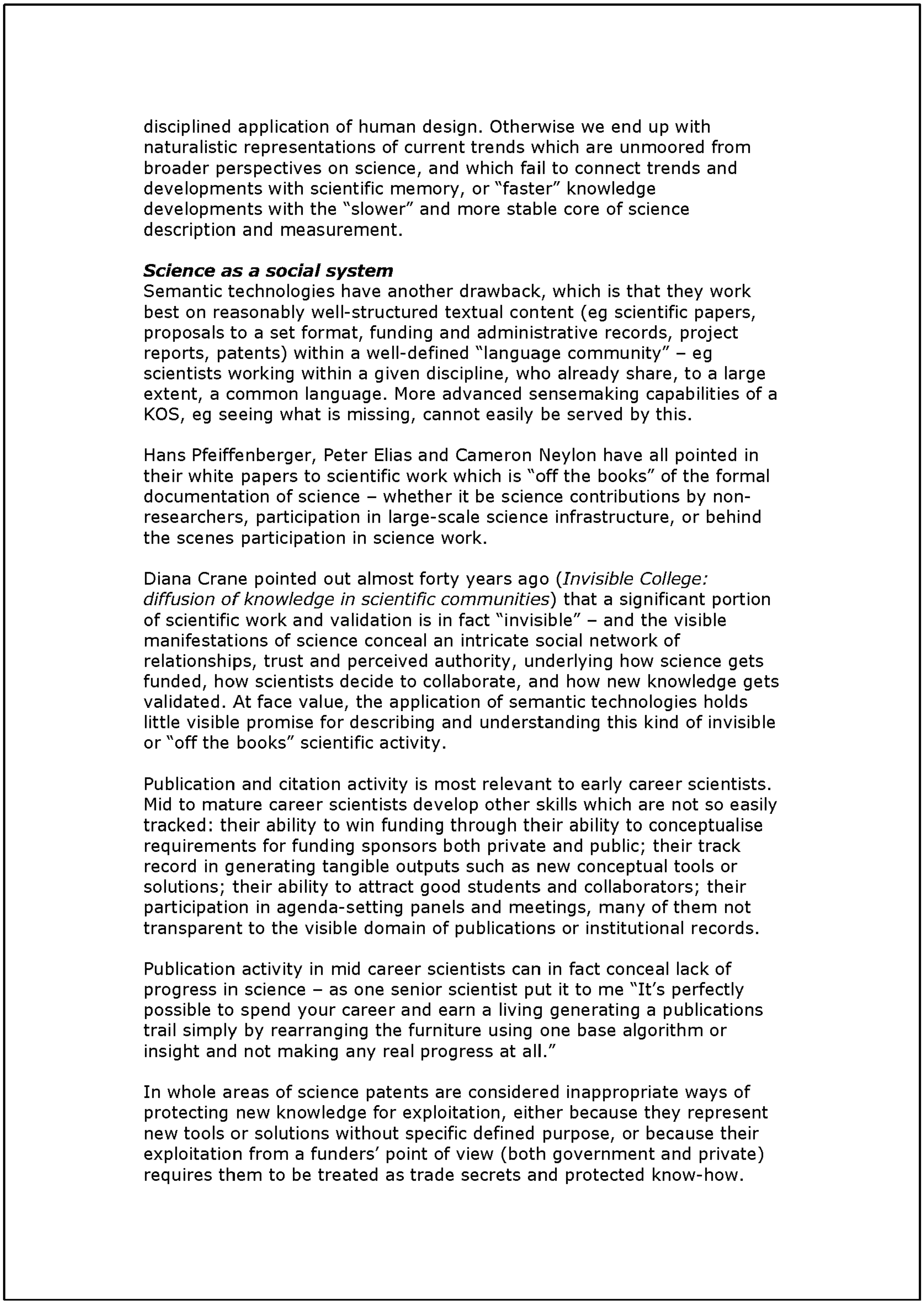
The typical managerial response to the human aversion to working with – and contributing to – a complex KOS in a disciplined and consistent way, is to use semantic technologies to analyse natural or semi-controlled language texts and to make inferences about topics and relationships between topics to feed the ontology-supported approach. These technologies have great potential for sidestepping human aversion to control and consistency, and they are also very powerful for identifying emerging trends in science – too much control suppresses new or variant language about science, and so suppresses signals of new science. Semantic technologies can also infer relationships between concepts, based on association patterns.

However, to perform the larger functions of coordination of language, meaningful connections and sensemaking in support of science, human intervention is required to judge and identify the most salient relationships, and to establish connections between domains as well as between past and future science language.

Principle 3: it is not sufficient to use semantic technology to describe science activity. This does not get at all the functions of a KOS. Linnaeus and Mendeleev had the impact they had, because they engaged in a work of design, not simply description.

In practice in today's world, the task is no longer within the grasp of gifted and determined individuals such as Linnaeus and Mendeleev. We require institutional interventions, in the form of development and maintenance of standardised vocabularies, taxonomies and ontologies, and the environments where they can be deployed.

Any KOS intended to meet the needs of understanding and progressing science will require some elements of designed structure and the



disciplined application of human design. Otherwise we end up with naturalistic representations of current trends which are unmoored from broader perspectives on science, and which fail to connect trends and developments with scientific memory, or "faster" knowledge developments with the "slower" and more stable core of science description and measurement.

Science as a social system

Semantic technologies have another drawback, which is that they work best on reasonably well-structured textual content (eg scientific papers, proposals to a set format, funding and administrative records, project reports, patents) within a well-defined "language community" – eg scientists working within a given discipline, who already share, to a large extent, a common language. More advanced sensemaking capabilities of a KOS, eg seeing what is missing, cannot easily be served by this.

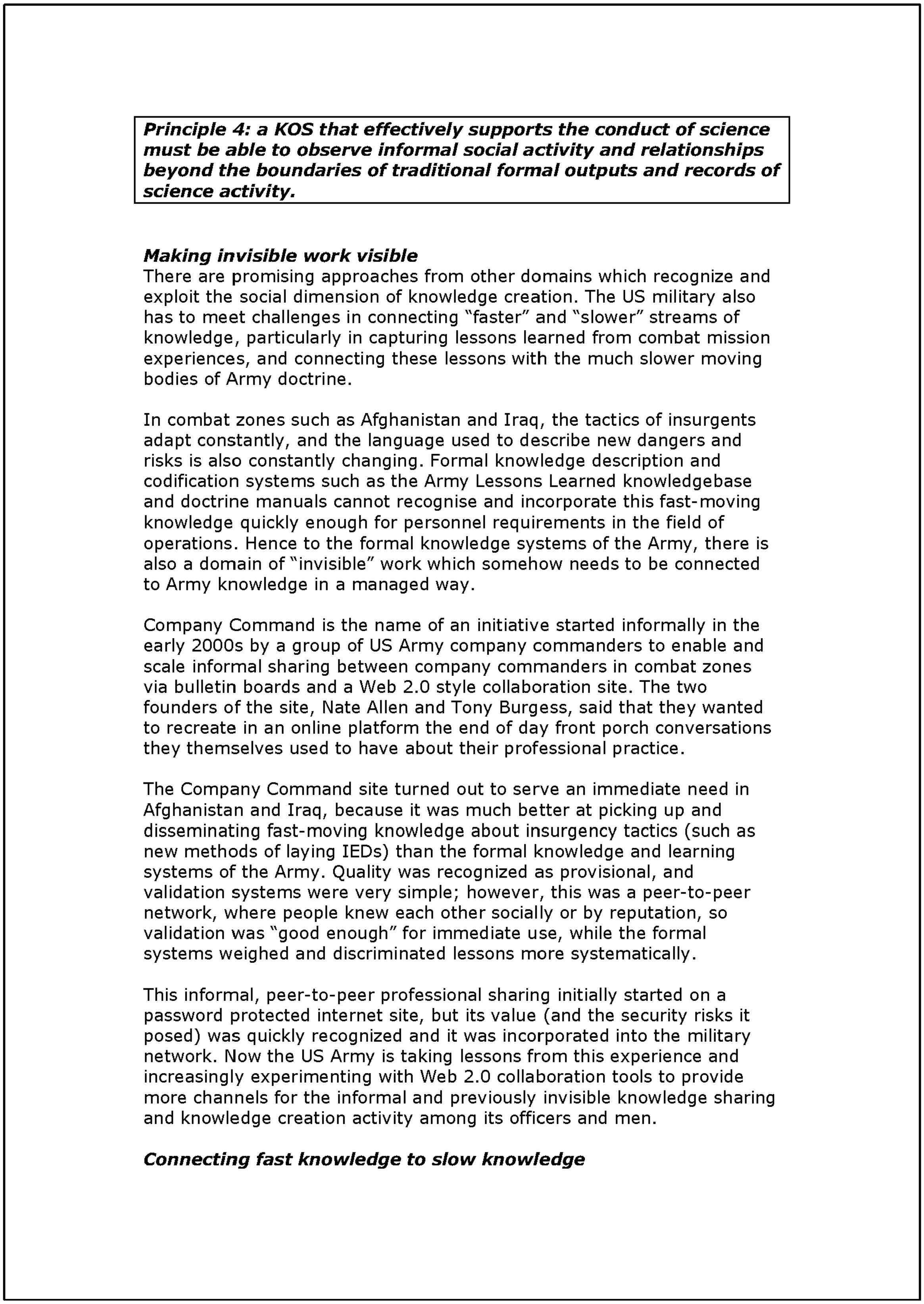
Hans Pfeiffenberger, Peter Elias and Cameron Neylon have all pointed in their white papers to scientific work which is "off the books" of the formal documentation of science – whether it be science contributions by non-researchers, participation in large-scale science infrastructure, or behind the scenes participation in science work.

Diana Crane pointed out almost forty years ago (*Invisible College: diffusion of knowledge in scientific communities*) that a significant portion of scientific work and validation is in fact "invisible" – and the visible manifestations of science conceal an intricate social network of relationships, trust and perceived authority, underlying how science gets funded, how scientists decide to collaborate, and how new knowledge gets validated. At face value, the application of semantic technologies holds little visible promise for describing and understanding this kind of invisible or "off the books" scientific activity.

Publication and citation activity is most relevant to early career scientists. Mid to mature career scientists develop other skills which are not so easily tracked: their ability to win funding through their ability to conceptualise requirements for funding sponsors both private and public; their track record in generating tangible outputs such as new conceptual tools or solutions; their ability to attract good students and collaborators; their participation in agenda-setting panels and meetings, many of them not transparent to the visible domain of publications or institutional records.

Publication activity in mid career scientists can in fact conceal lack of progress in science – as one senior scientist put it to me "It's perfectly possible to spend your career and earn a living generating a publications trail simply by rearranging the furniture using one base algorithm or insight and not making any real progress at all."

In whole areas of science patents are considered inappropriate ways of protecting new knowledge for exploitation, either because they represent new tools or solutions without specific defined purpose, or because their exploitation from a funders' point of view (both government and private) requires them to be treated as trade secrets and protected know-how.



Principle 4: a KOS that effectively supports the conduct of science must be able to observe informal social activity and relationships beyond the boundaries of traditional formal outputs and records of science activity.

Making invisible work visible

There are promising approaches from other domains which recognize and exploit the social dimension of knowledge creation. The US military also has to meet challenges in connecting "faster" and "slower" streams of knowledge, particularly in capturing lessons learned from combat mission experiences, and connecting these lessons with the much slower moving bodies of Army doctrine.

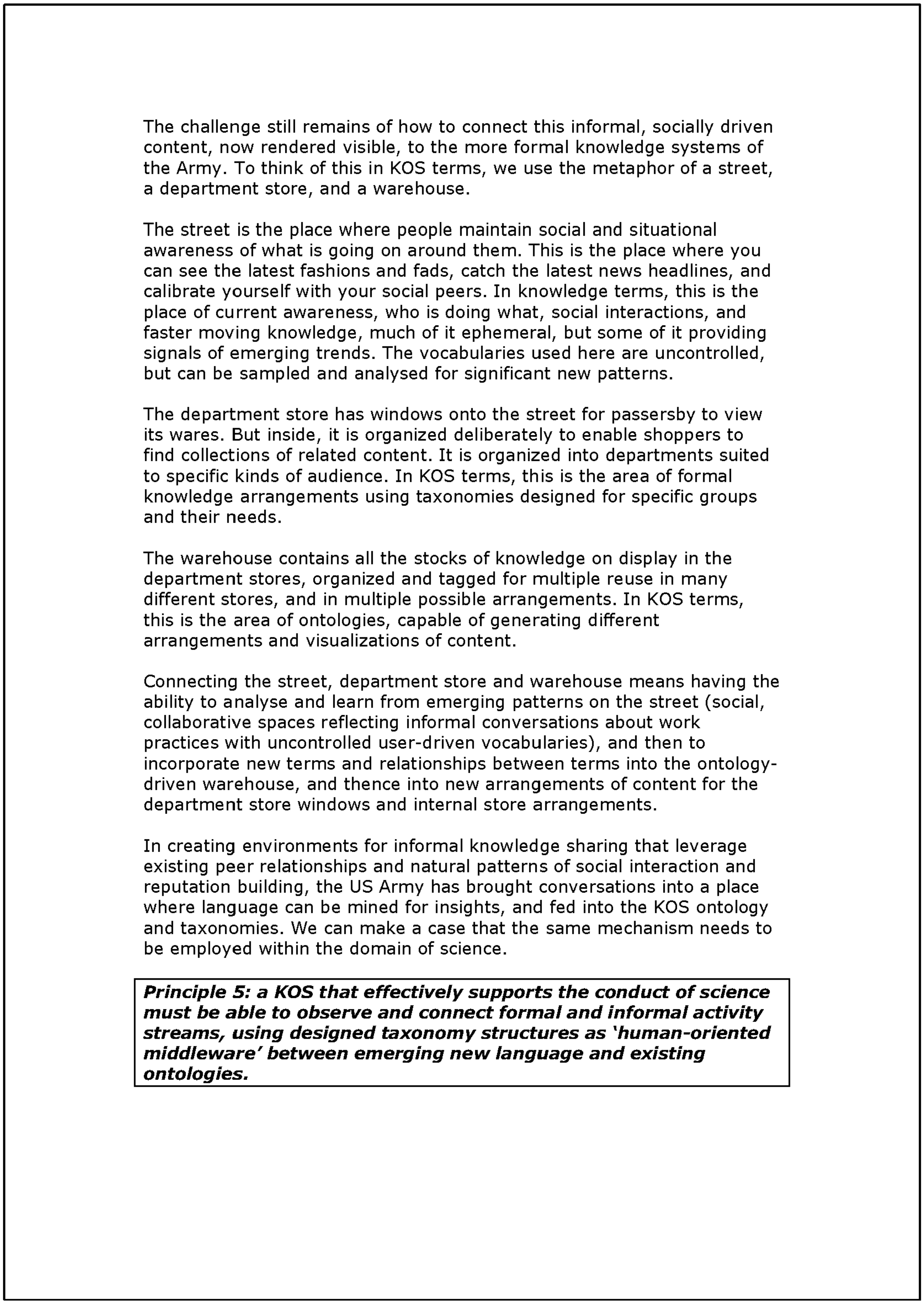
In combat zones such as Afghanistan and Iraq, the tactics of insurgents adapt constantly, and the language used to describe new dangers and risks is also constantly changing. Formal knowledge description and codification systems such as the Army Lessons Learned knowledgebase and doctrine manuals cannot recognise and incorporate this fast-moving knowledge quickly enough for personnel requirements in the field of operations. Hence to the formal knowledge systems of the Army, there is also a domain of "invisible" work which somehow needs to be connected to Army knowledge in a managed way.

Company Command is the name of an initiative started informally in the early 2000s by a group of US Army company commanders to enable and scale informal sharing between company commanders in combat zones via bulletin boards and a Web 2.0 style collaboration site. The two founders of the site, Nate Allen and Tony Burgess, said that they wanted to recreate in an online platform the end of day front porch conversations they themselves used to have about their professional practice.

The Company Command site turned out to serve an immediate need in Afghanistan and Iraq, because it was much better at picking up and disseminating fast-moving knowledge about insurgency tactics (such as new methods of laying IEDs) than the formal knowledge and learning systems of the Army. Quality was recognized as provisional, and validation systems were very simple; however, this was a peer-to-peer network, where people knew each other socially or by reputation, so validation was "good enough" for immediate use, while the formal systems weighed and discriminated lessons more systematically.

This informal, peer-to-peer professional sharing initially started on a password protected internet site, but its value (and the security risks it posed) was quickly recognized and it was incorporated into the military network. Now the US Army is taking lessons from this experience and increasingly experimenting with Web 2.0 collaboration tools to provide more channels for the informal and previously invisible knowledge sharing and knowledge creation activity among its officers and men.

Connecting fast knowledge to slow knowledge



The challenge still remains of how to connect this informal, socially driven content, now rendered visible, to the more formal knowledge systems of the Army. To think of this in KOS terms, we use the metaphor of a street, a department store, and a warehouse.

The street is the place where people maintain social and situational awareness of what is going on around them. This is the place where you can see the latest fashions and fads, catch the latest news headlines, and calibrate yourself with your social peers. In knowledge terms, this is the place of current awareness, who is doing what, social interactions, and faster moving knowledge, much of it ephemeral, but some of it providing signals of emerging trends. The vocabularies used here are uncontrolled, but can be sampled and analysed for significant new patterns.

The department store has windows onto the street for passersby to view its wares. But inside, it is organized deliberately to enable shoppers to find collections of related content. It is organized into departments suited to specific kinds of audience. In KOS terms, this is the area of formal knowledge arrangements using taxonomies designed for specific groups and their needs.

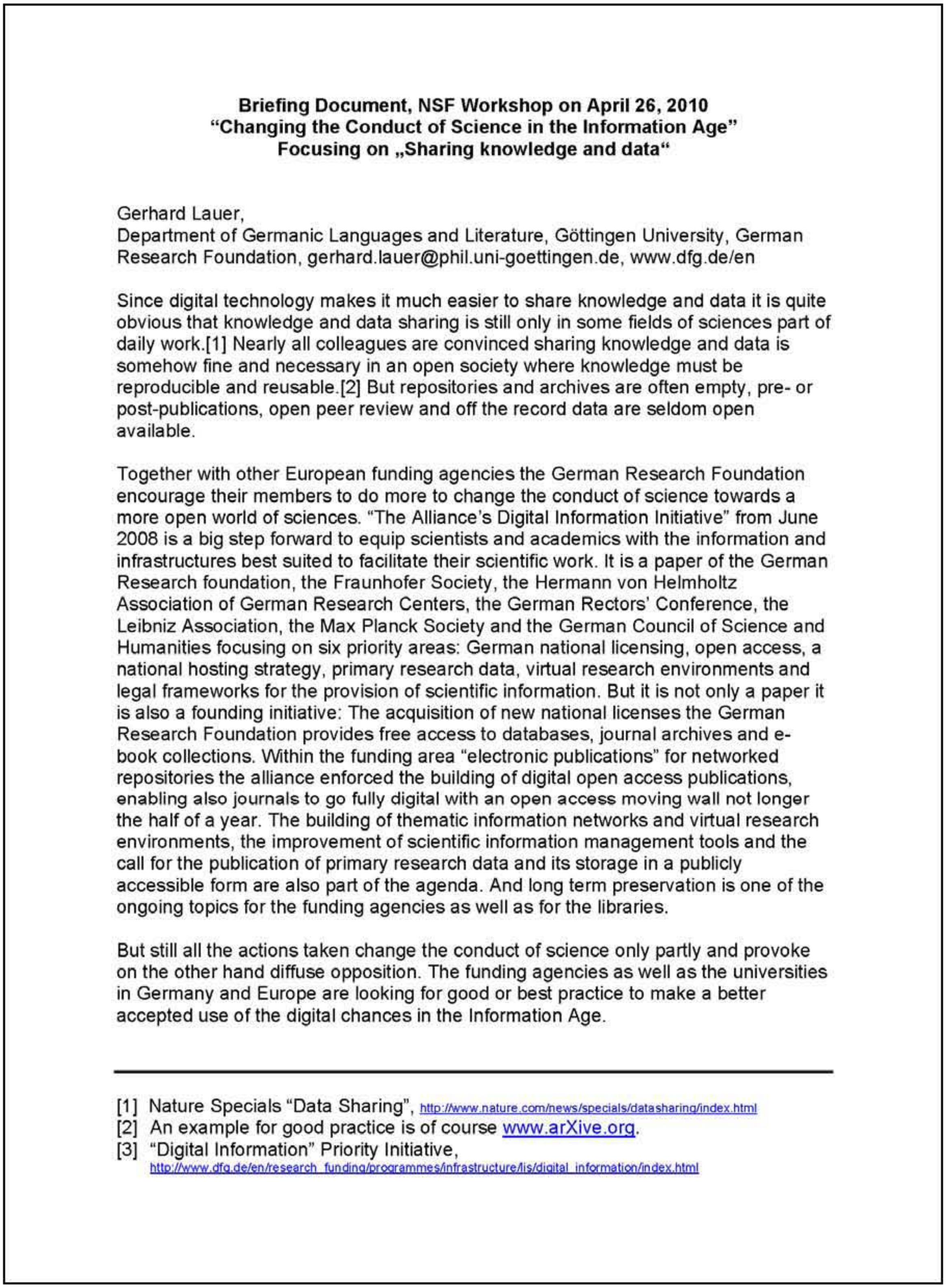
The warehouse contains all the stocks of knowledge on display in the department stores, organized and tagged for multiple reuse in many different stores, and in multiple possible arrangements. In KOS terms, this is the area of ontologies, capable of generating different arrangements and visualizations of content.

Connecting the street, department store and warehouse means having the ability to analyse and learn from emerging patterns on the street (social, collaborative spaces reflecting informal conversations about work practices with uncontrolled user-driven vocabularies), and then to incorporate new terms and relationships between terms into the ontology-driven warehouse, and thence into new arrangements of content for the department store windows and internal store arrangements.

In creating environments for informal knowledge sharing that leverage existing peer relationships and natural patterns of social interaction and reputation building, the US Army has brought conversations into a place where language can be mined for insights, and fed into the KOS ontology and taxonomies. We can make a case that the same mechanism needs to be employed within the domain of science.

Principle 5: a KOS that effectively supports the conduct of science must be able to observe and connect formal and informal activity streams, using designed taxonomy structures as 'human-oriented middleware' between emerging new language and existing ontologies.

Lauer, Gerhard, "Changing the Conduct of Science in the Information Age: Focusing on Sharing Knowledge and Data"



Briefing Document, NSF Workshop on April 26, 2010
"Changing the Conduct of Science in the Information Age"
Focusing on „Sharing knowledge and data"

Gerhard Lauer,
Department of Germanic Languages and Literature, Göttingen University, German Research Foundation, gerhard.lauer@phil.uni-goettingen.de, www.dfg.de/en

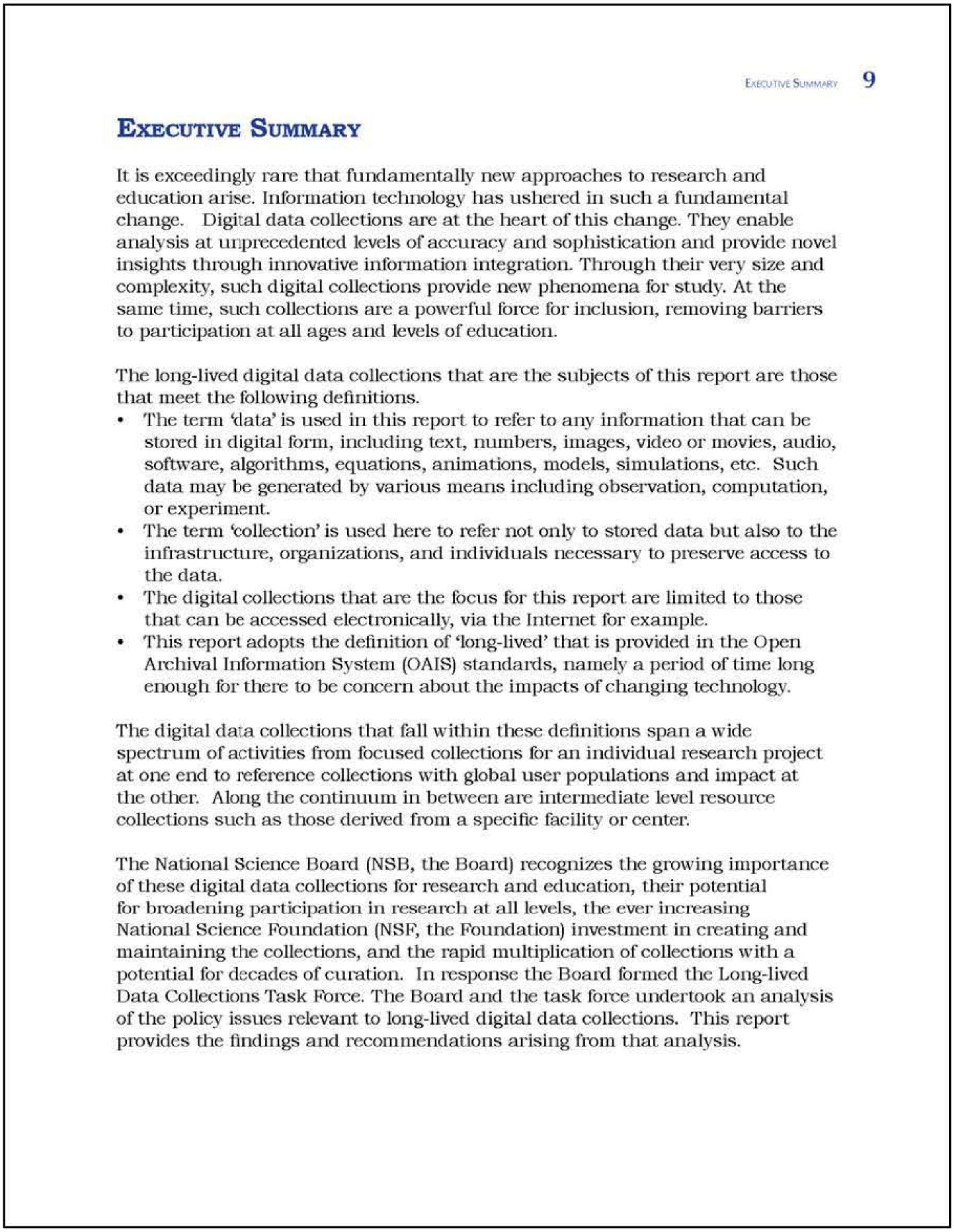
Since digital technology makes it much easier to share knowledge and data it is quite obvious that knowledge and data sharing is still only in some fields of sciences part of daily work.[1] Nearly all colleagues are convinced sharing knowledge and data is somehow fine and necessary in an open society where knowledge must be reproducible and reusable.[2] But repositories and archives are often empty, pre- or post-publications, open peer review and off the record data are seldom open available.

Together with other European funding agencies the German Research Foundation encourage their members to do more to change the conduct of science towards a more open world of sciences. "The Alliance's Digital Information Initiative" from June 2008 is a big step forward to equip scientists and academics with the information and infrastructures best suited to facilitate their scientific work. It is a paper of the German Research foundation, the Fraunhofer Society, the Hermann von Helmholtz Association of German Research Centers, the German Rectors' Conference, the Leibniz Association, the Max Planck Society and the German Council of Science and Humanities focusing on six priority areas: German national licensing, open access, a national hosting strategy, primary research data, virtual research environments and legal frameworks for the provision of scientific information. But it is not only a paper it is also a founding initiative: The acquisition of new national licenses the German Research Foundation provides free access to databases, journal archives and e-book collections. Within the funding area "electronic publications" for networked repositories the alliance enforced the building of digital open access publications, enabling also journals to go fully digital with an open access moving wall not longer the half of a year. The building of thematic information networks and virtual research environments, the improvement of scientific information management tools and the call for the publication of primary research data and its storage in a publicly accessible form are also part of the agenda. And long term preservation is one of the ongoing topics for the funding agencies as well as for the libraries.

But still all the actions taken change the conduct of science only partly and provoke on the other hand diffuse opposition. The funding agencies as well as the universities in Germany and Europe are looking for good or best practice to make a better accepted use of the digital chances in the Information Age.

[1] Nature Specials "Data Sharing", http://www.nature.com/news/specials/datasharing/index.html
[2] An example for good practice is of course www.arXive.org.
[3] "Digital Information" Priority Initiative, http://www.dfg.de/en/research_funding/programmes/infrastructure/lis/digital_information/index.html

National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century"



Executive Summary

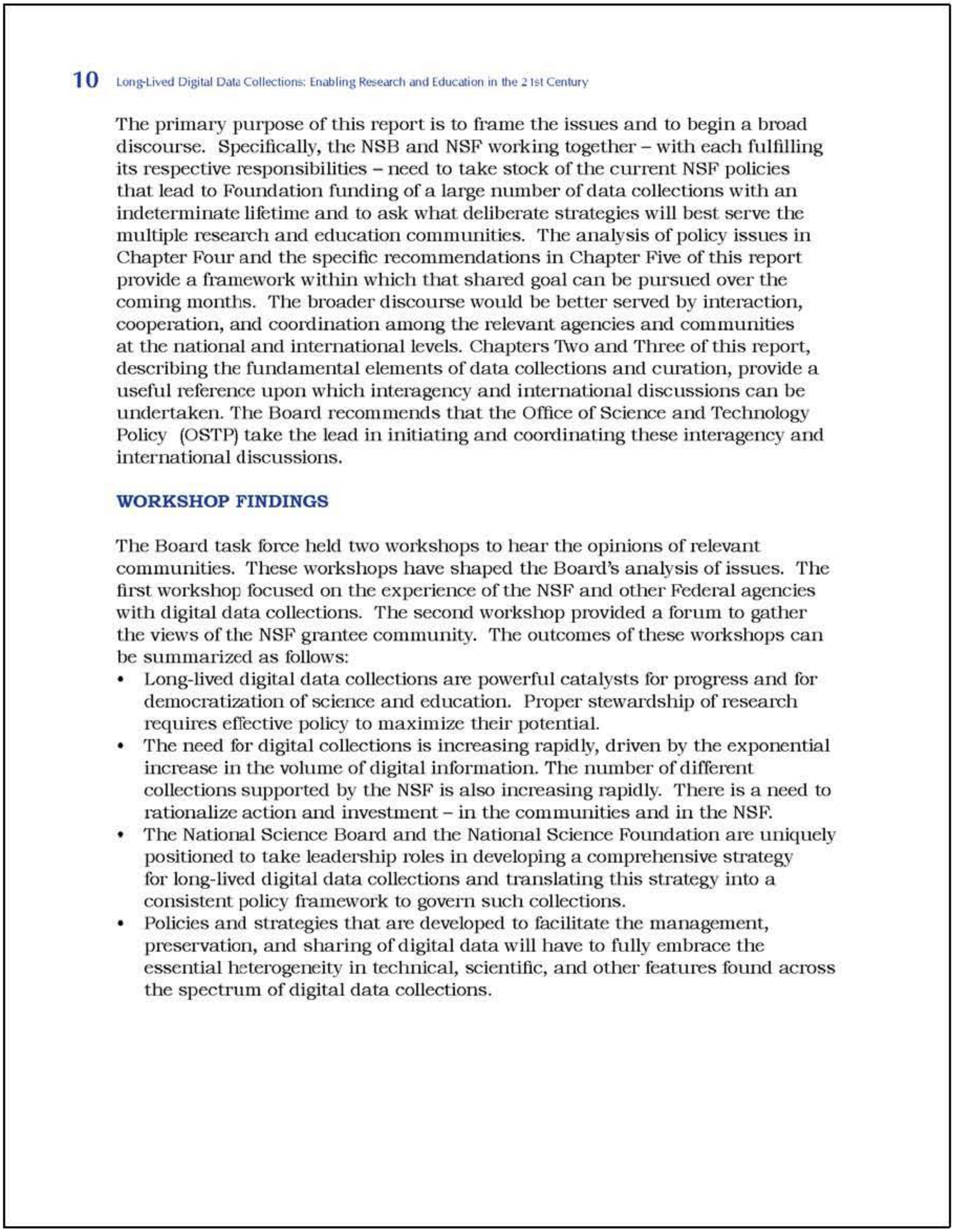
It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.

The long-lived digital data collections that are the subjects of this report are those that meet the following definitions.

- The term 'data' is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment.
- The term 'collection' is used here to refer not only to stored data but also to the infrastructure, organizations, and individuals necessary to preserve access to the data.
- The digital collections that are the focus for this report are limited to those that can be accessed electronically, via the Internet for example.
- This report adopts the definition of 'long-lived' that is provided in the Open Archival Information System (OAIS) standards, namely a period of time long enough for there to be concern about the impacts of changing technology.

The digital data collections that fall within these definitions span a wide spectrum of activities from focused collections for an individual research project at one end to reference collections with global user populations and impact at the other. Along the continuum in between are intermediate level resource collections such as those derived from a specific facility or center.

The National Science Board (NSB, the Board) recognizes the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, the ever increasing National Science Foundation (NSF, the Foundation) investment in creating and maintaining the collections, and the rapid multiplication of collections with a potential for decades of curation. In response the Board formed the Long-lived Data Collections Task Force. The Board and the task force undertook an analysis of the policy issues relevant to long-lived digital data collections. This report provides the findings and recommendations arising from that analysis.

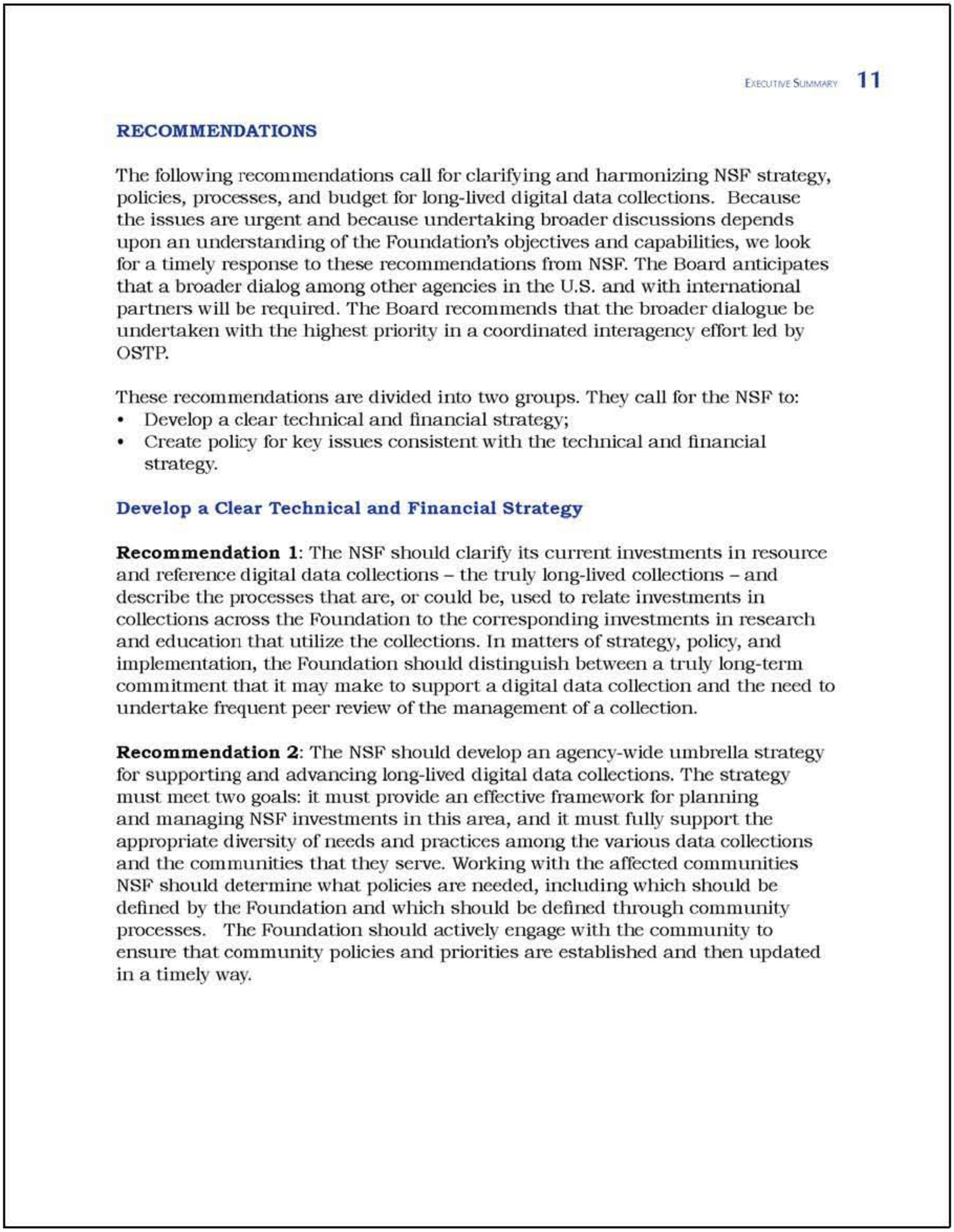


The primary purpose of this report is to frame the issues and to begin a broad discourse. Specifically, the NSB and NSF working together – with each fulfilling its respective responsibilities – need to take stock of the current NSF policies that lead to Foundation funding of a large number of data collections with an indeterminate lifetime and to ask what deliberate strategies will best serve the multiple research and education communities. The analysis of policy issues in Chapter Four and the specific recommendations in Chapter Five of this report provide a framework within which that shared goal can be pursued over the coming months. The broader discourse would be better served by interaction, cooperation, and coordination among the relevant agencies and communities at the national and international levels. Chapters Two and Three of this report, describing the fundamental elements of data collections and curation, provide a useful reference upon which interagency and international discussions can be undertaken. The Board recommends that the Office of Science and Technology Policy (OSTP) take the lead in initiating and coordinating these interagency and international discussions.

WORKSHOP FINDINGS

The Board task force held two workshops to hear the opinions of relevant communities. These workshops have shaped the Board's analysis of issues. The first workshop focused on the experience of the NSF and other Federal agencies with digital data collections. The second workshop provided a forum to gather the views of the NSF grantee community. The outcomes of these workshops can be summarized as follows:

- Long-lived digital data collections are powerful catalysts for progress and for democratization of science and education. Proper stewardship of research requires effective policy to maximize their potential.
- The need for digital collections is increasing rapidly, driven by the exponential increase in the volume of digital information. The number of different collections supported by the NSF is also increasing rapidly. There is a need to rationalize action and investment – in the communities and in the NSF.
- The National Science Board and the National Science Foundation are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections and translating this strategy into a consistent policy framework to govern such collections.
- Policies and strategies that are developed to facilitate the management, preservation, and sharing of digital data will have to fully embrace the essential heterogeneity in technical, scientific, and other features found across the spectrum of digital data collections.



RECOMMENDATIONS

The following recommendations call for clarifying and harmonizing NSF strategy, policies, processes, and budget for long-lived digital data collections. Because the issues are urgent and because undertaking broader discussions depends upon an understanding of the Foundation's objectives and capabilities, we look for a timely response to these recommendations from NSF. The Board anticipates that a broader dialog among other agencies in the U.S. and with international partners will be required. The Board recommends that the broader dialogue be undertaken with the highest priority in a coordinated interagency effort led by OSTP.

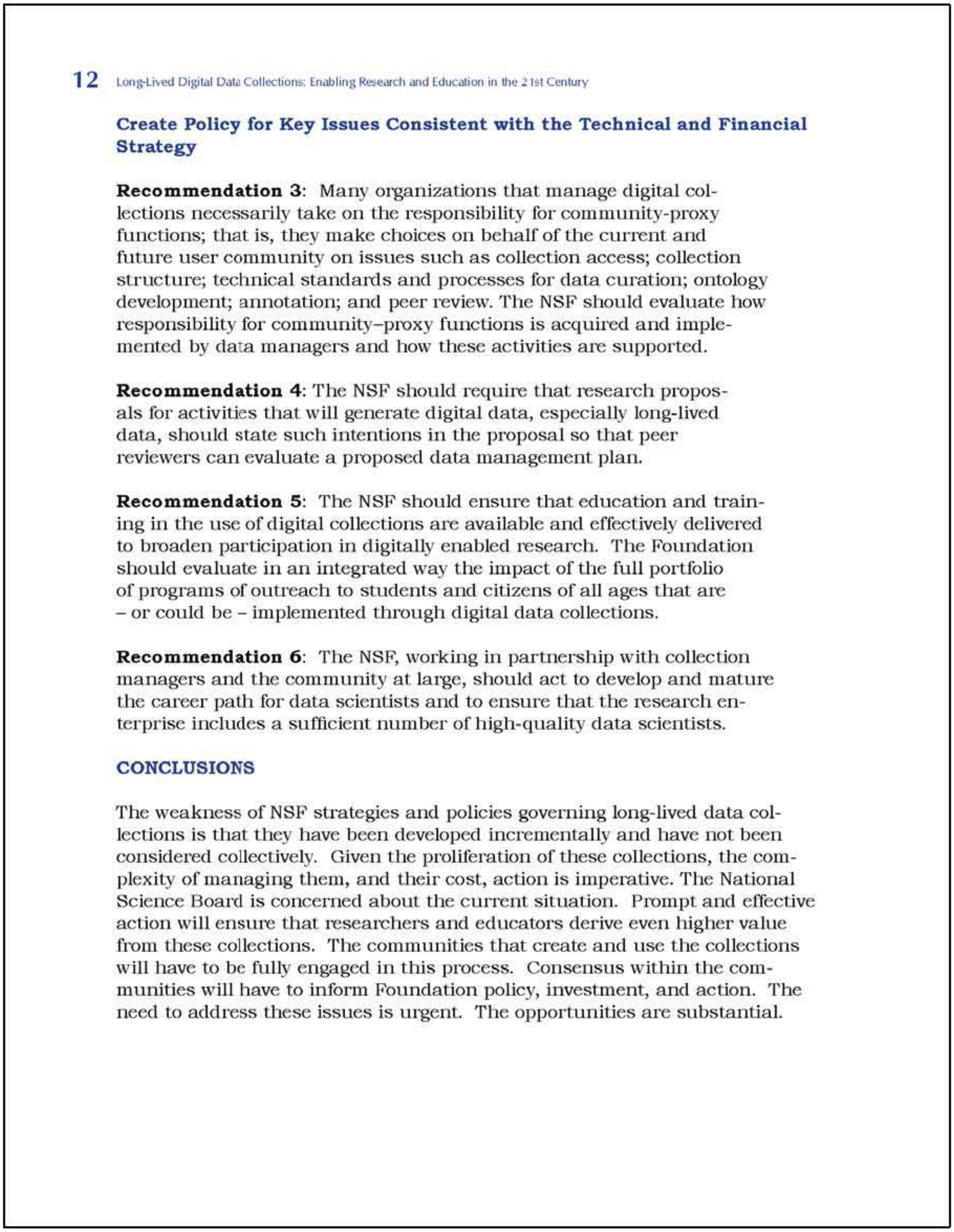
These recommendations are divided into two groups. They call for the NSF to:

- Develop a clear technical and financial strategy;
- Create policy for key issues consistent with the technical and financial strategy.

Develop a Clear Technical and Financial Strategy

Recommendation 1: The NSF should clarify its current investments in resource and reference digital data collections – the truly long-lived collections – and describe the processes that are, or could be, used to relate investments in collections across the Foundation to the corresponding investments in research and education that utilize the collections. In matters of strategy, policy, and implementation, the Foundation should distinguish between a truly long-term commitment that it may make to support a digital data collection and the need to undertake frequent peer review of the management of a collection.

Recommendation 2: The NSF should develop an agency-wide umbrella strategy for supporting and advancing long-lived digital data collections. The strategy must meet two goals: it must provide an effective framework for planning and managing NSF investments in this area, and it must fully support the appropriate diversity of needs and practices among the various data collections and the communities that they serve. Working with the affected communities NSF should determine what policies are needed, including which should be defined by the Foundation and which should be defined through community processes. The Foundation should actively engage with the community to ensure that community policies and priorities are established and then updated in a timely way.



12 Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

Create Policy for Key Issues Consistent with the Technical and Financial Strategy

Recommendation 3: Many organizations that manage digital collections necessarily take on the responsibility for community-proxy functions; that is, they make choices on behalf of the current and future user community on issues such as collection access; collection structure; technical standards and processes for data curation; ontology development; annotation; and peer review. The NSF should evaluate how responsibility for community–proxy functions is acquired and implemented by data managers and how these activities are supported.

Recommendation 4: The NSF should require that research proposals for activities that will generate digital data, especially long-lived data, should state such intentions in the proposal so that peer reviewers can evaluate a proposed data management plan.

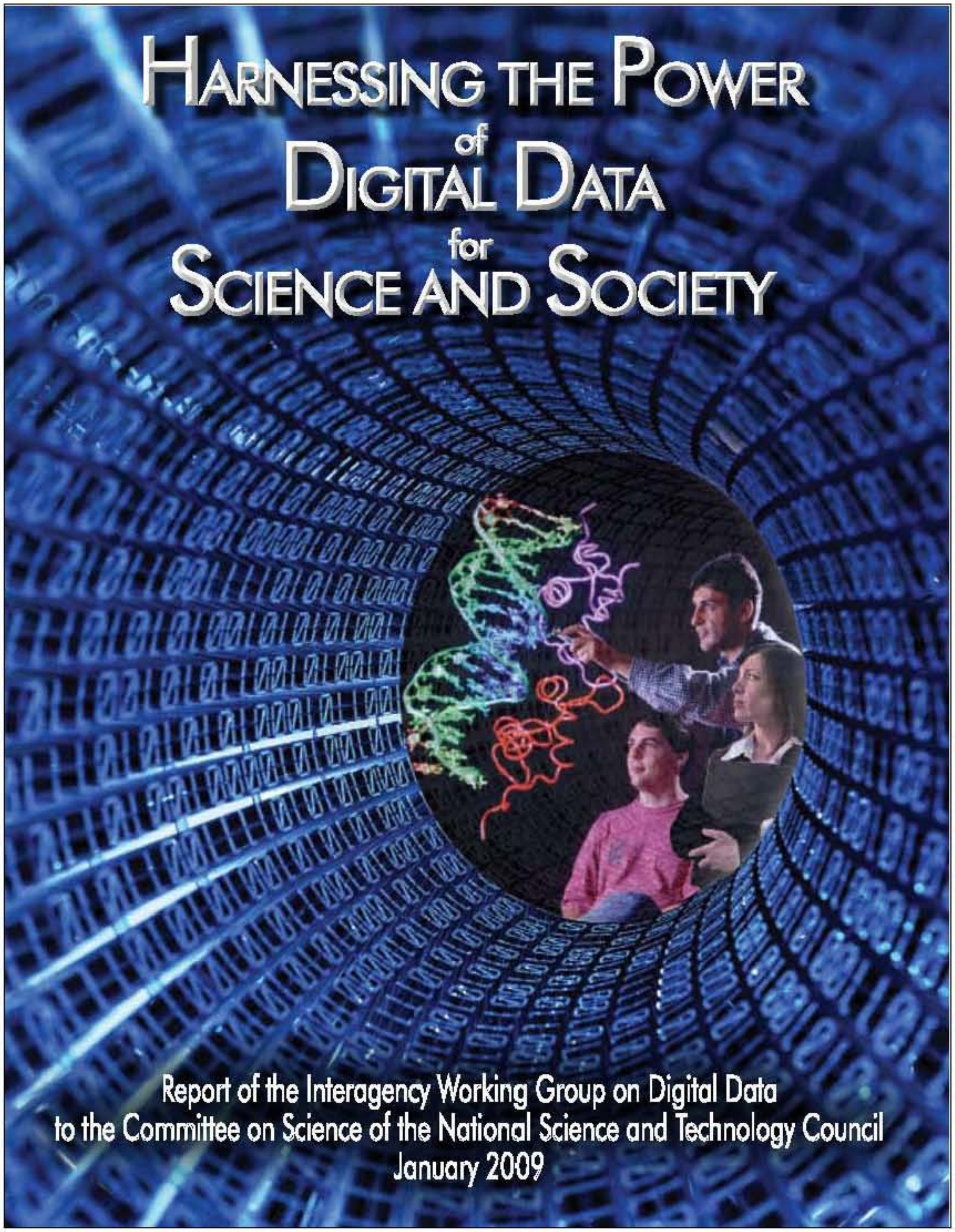
Recommendation 5: The NSF should ensure that education and training in the use of digital collections are available and effectively delivered to broaden participation in digitally enabled research. The Foundation should evaluate in an integrated way the impact of the full portfolio of programs of outreach to students and citizens of all ages that are – or could be – implemented through digital data collections.

Recommendation 6: The NSF, working in partnership with collection managers and the community at large, should act to develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists.

CONCLUSIONS

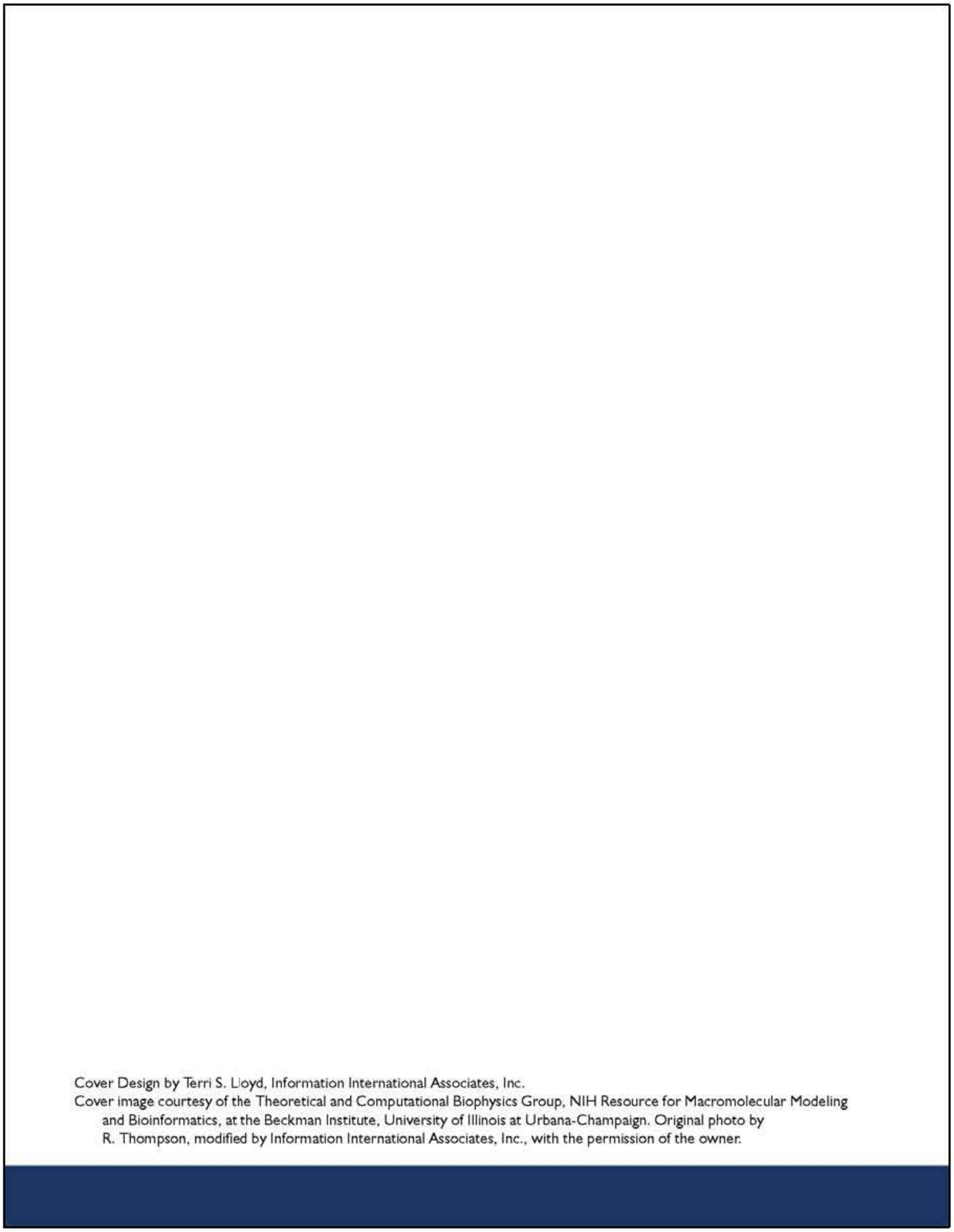
The weakness of NSF strategies and policies governing long-lived data collections is that they have been developed incrementally and have not been considered collectively. Given the proliferation of these collections, the complexity of managing them, and their cost, action is imperative. The National Science Board is concerned about the current situation. Prompt and effective action will ensure that researchers and educators derive even higher value from these collections. The communities that create and use the collections will have to be fully engaged in this process. Consensus within the communities will have to inform Foundation policy, investment, and action. The need to address these issues is urgent. The opportunities are substantial.

Office of Science and Technology Policy, "Harnessing the Power of Digital Data for Science and Society"

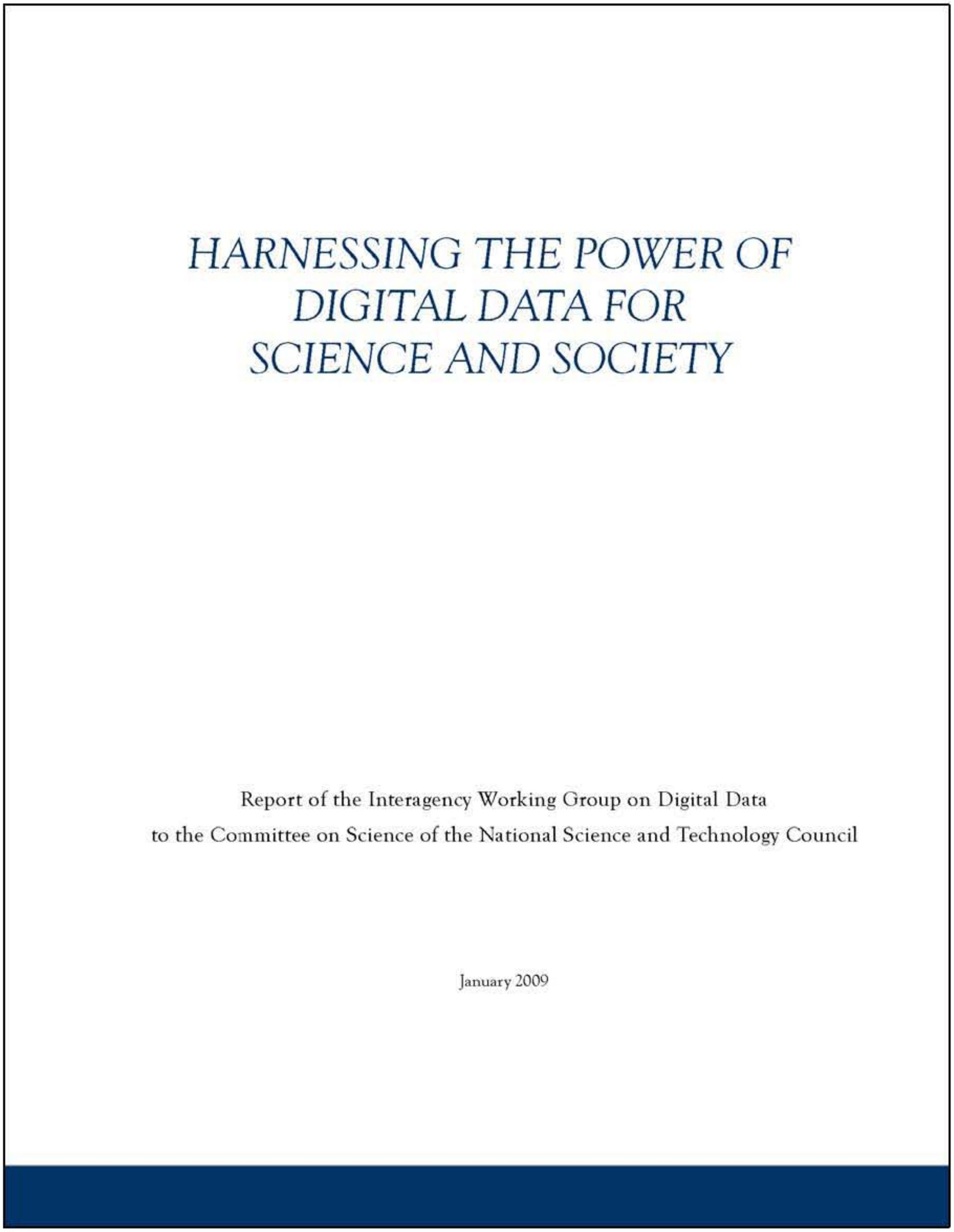


HARNESSING THE POWER of DIGITAL DATA for SCIENCE AND SOCIETY

Report of the Interagency Working Group on Digital Data
to the Committee on Science of the National Science and Technology Council
January 2009



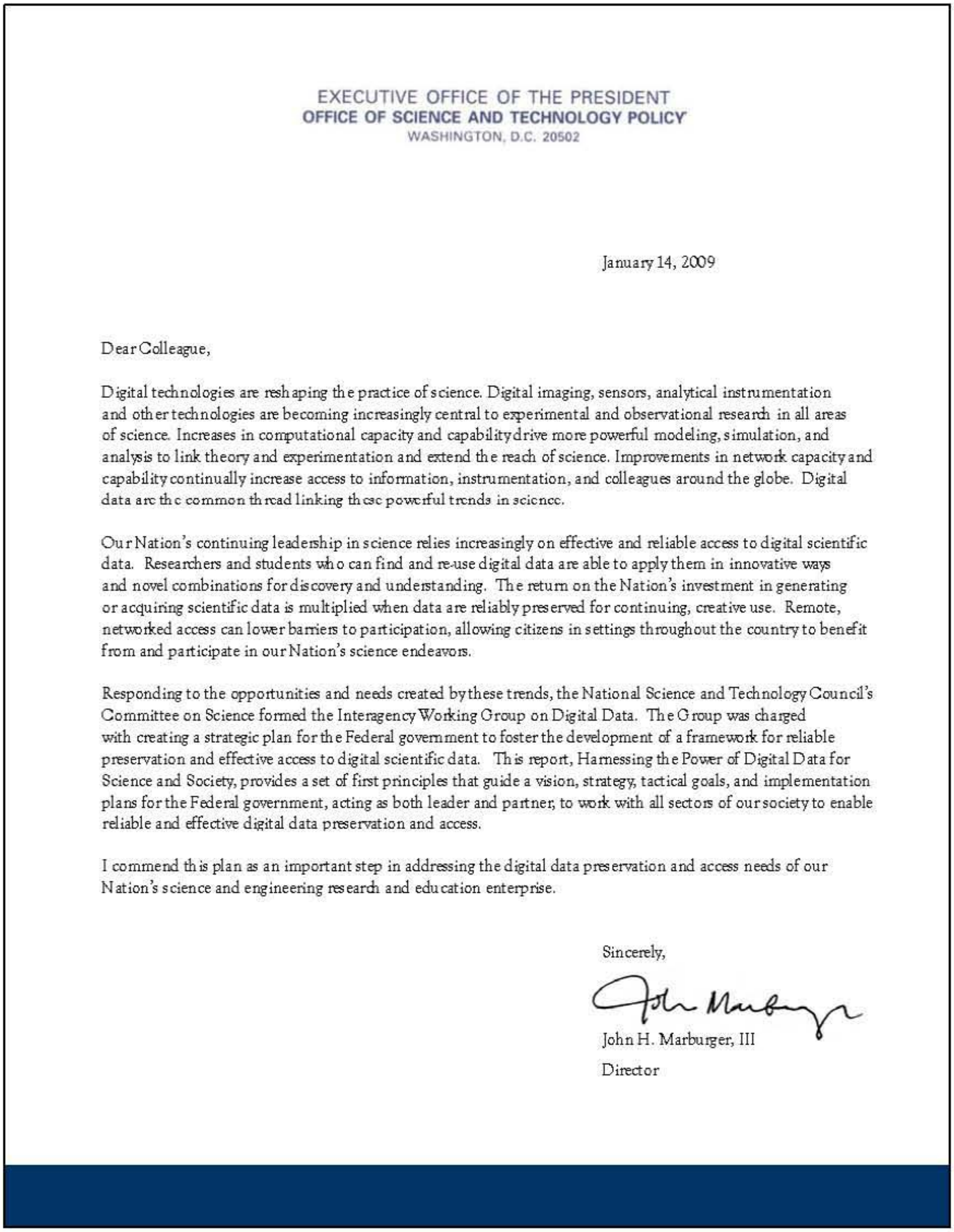
Cover Design by Terri S. Lloyd, Information International Associates, Inc.
Cover image courtesy of the Theoretical and Computational Biophysics Group, NIH Resource for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign. Original photo by R. Thompson, modified by Information International Associates, Inc., with the permission of the owner.



HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY

Report of the Interagency Working Group on Digital Data
to the Committee on Science of the National Science and Technology Council

January 2009



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

January 14, 2009

Dear Colleague,

Digital technologies are reshaping the practice of science. Digital imaging, sensors, analytical instrumentation and other technologies are becoming increasingly central to experimental and observational research in all areas of science. Increases in computational capacity and capability drive more powerful modeling, simulation, and analysis to link theory and experimentation and extend the reach of science. Improvements in network capacity and capability continually increase access to information, instrumentation, and colleagues around the globe. Digital data are the common thread linking these powerful trends in science.

Our Nation's continuing leadership in science relies increasingly on effective and reliable access to digital scientific data. Researchers and students who can find and re-use digital data are able to apply them in innovative ways and novel combinations for discovery and understanding. The return on the Nation's investment in generating or acquiring scientific data is multiplied when data are reliably preserved for continuing, creative use. Remote, networked access can lower barriers to participation, allowing citizens in settings throughout the country to benefit from and participate in our Nation's science endeavors.

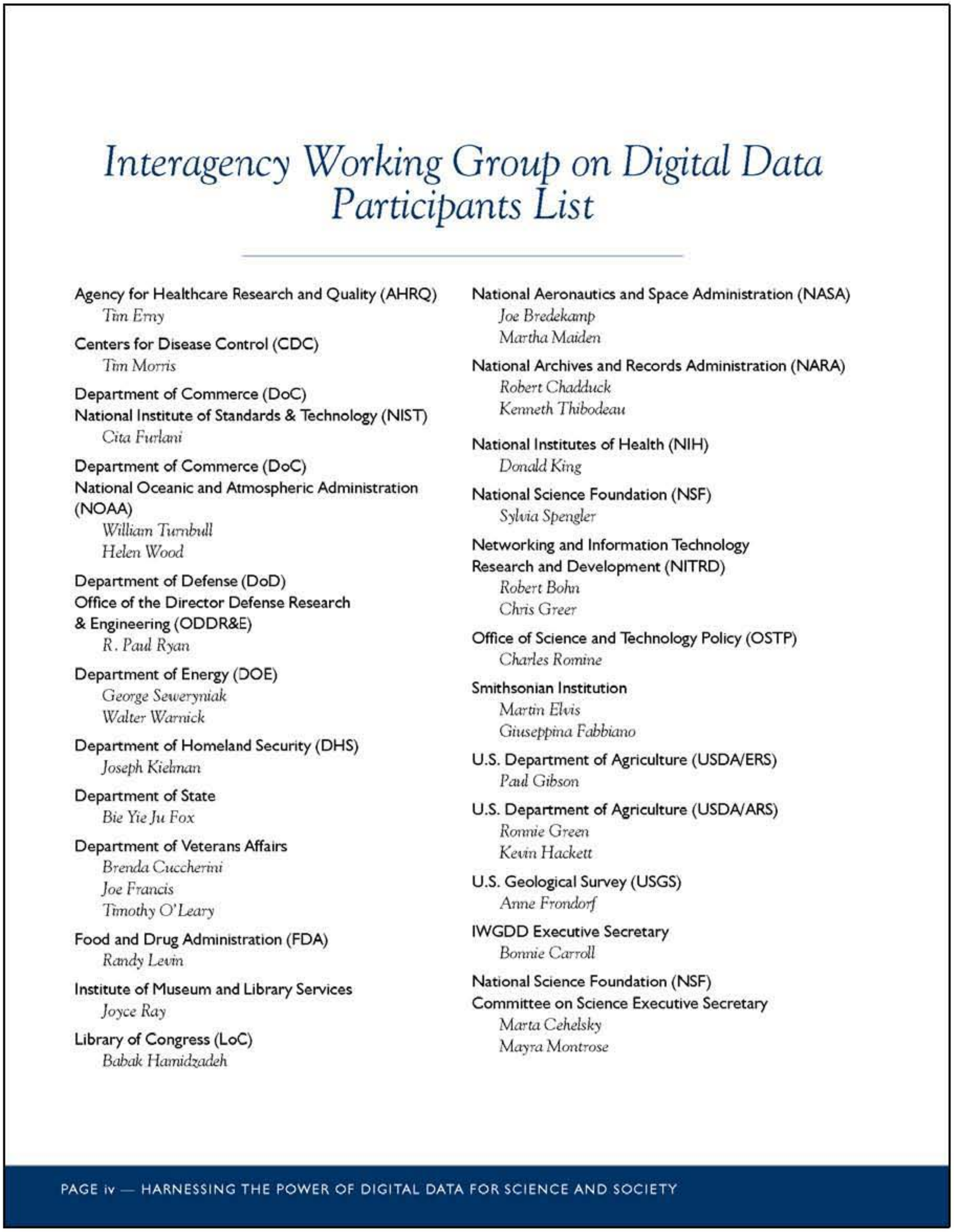
Responding to the opportunities and needs created by these trends, the National Science and Technology Council's Committee on Science formed the Interagency Working Group on Digital Data. The Group was charged with creating a strategic plan for the Federal government to foster the development of a framework for reliable preservation and effective access to digital scientific data. This report, Harnessing the Power of Digital Data for Science and Society, provides a set of first principles that guide a vision, strategy, tactical goals, and implementation plans for the Federal government, acting as both leader and partner, to work with all sectors of our society to enable reliable and effective digital data preservation and access.

I commend this plan as an important step in addressing the digital data preservation and access needs of our Nation's science and engineering research and education enterprise.

Sincerely,

John Marburger

John H. Marburger, III
Director



Interagency Working Group on Digital Data Participants List

Agency for Healthcare Research and Quality (AHRQ)
Tim Erny

Centers for Disease Control (CDC)
Tim Morris

Department of Commerce (DoC)
National Institute of Standards & Technology (NIST)
Cita Furlani

Department of Commerce (DoC)
National Oceanic and Atmospheric Administration (NOAA)
William Turnbull
Helen Wood

Department of Defense (DoD)
Office of the Director Defense Research & Engineering (ODDR&E)
R. Paul Ryan

Department of Energy (DOE)
George Seweryniak
Walter Warnick

Department of Homeland Security (DHS)
Joseph Kielman

Department of State
Bie Yie Ju Fox

Department of Veterans Affairs
Brenda Cuccherini
Joe Francis
Timothy O'Leary

Food and Drug Administration (FDA)
Randy Levin

Institute of Museum and Library Services
Joyce Ray

Library of Congress (LoC)
Babak Hamidzadeh

National Aeronautics and Space Administration (NASA)
Joe Bredekamp
Martha Maiden

National Archives and Records Administration (NARA)
Robert Chadduck
Kenneth Thibodeau

National Institutes of Health (NIH)
Donald King

National Science Foundation (NSF)
Sylvia Spengler

Networking and Information Technology Research and Development (NITRD)
Robert Bohn
Chris Greer

Office of Science and Technology Policy (OSTP)
Charles Romine

Smithsonian Institution
Martin Elvis
Giuseppina Fabbiano

U.S. Department of Agriculture (USDA/ERS)
Paul Gibson

U.S. Department of Agriculture (USDA/ARS)
Ronnie Green
Kevin Hackett

U.S. Geological Survey (USGS)
Anne Frondorf

IWGDD Executive Secretary
Bonnie Carroll

National Science Foundation (NSF)
Committee on Science Executive Secretary
Marta Cehelsky
Mayra Montrose

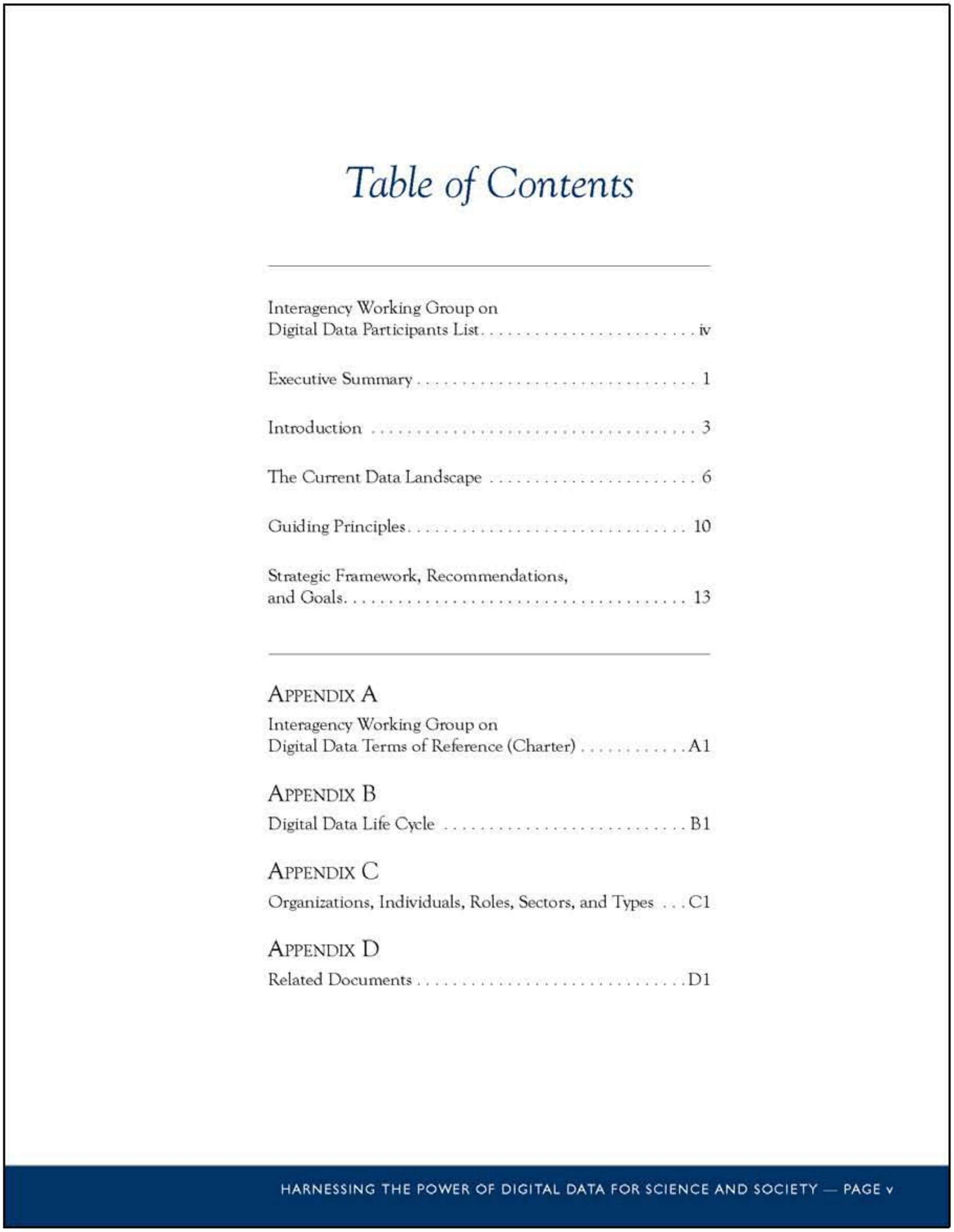


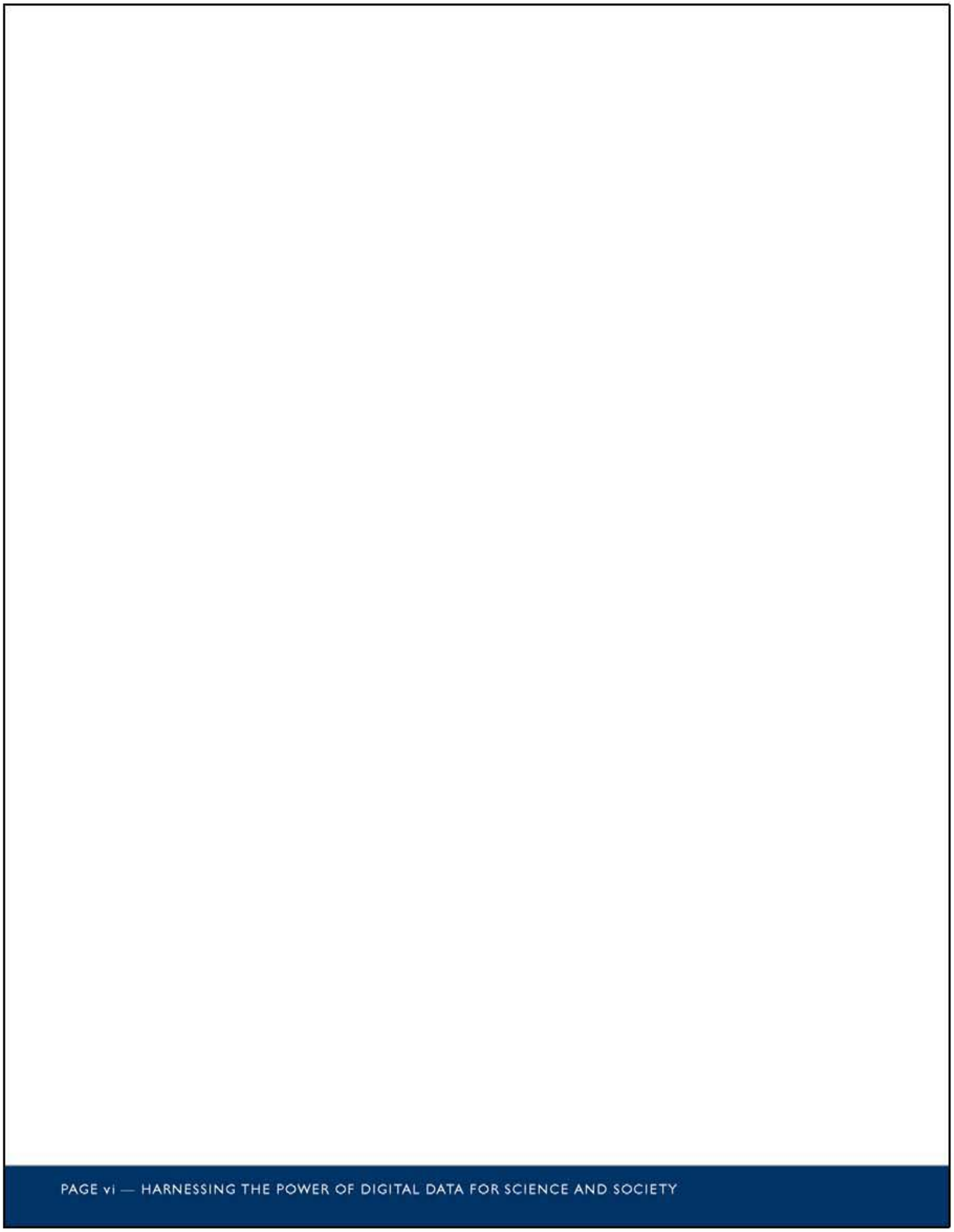
Table of Contents

Appendix A

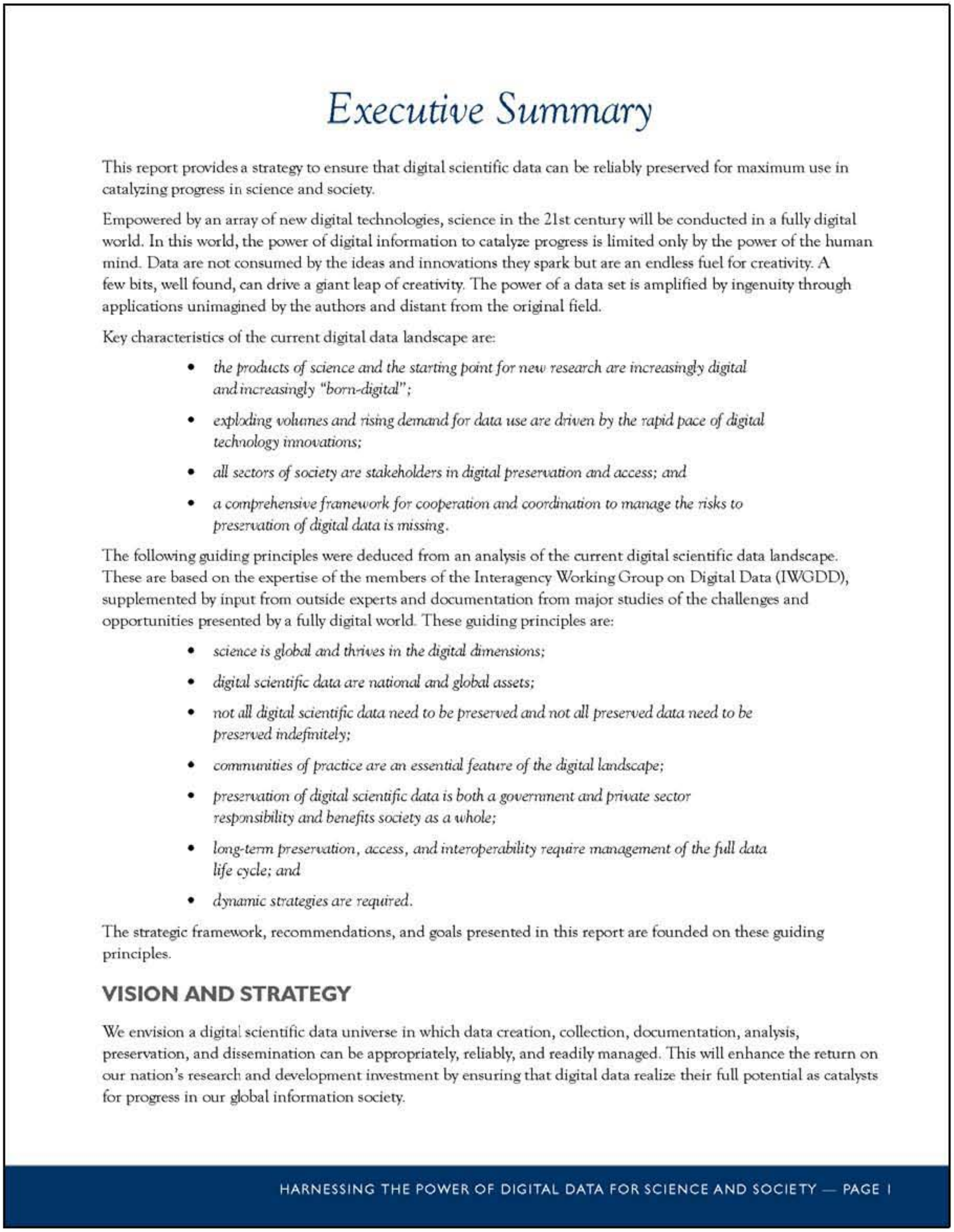
Appendix B

Appendix C

Appendix D



PAGE vi — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



Executive Summary

This report provides a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society.

Empowered by an array of new digital technologies, science in the 21st century will be conducted in a fully digital world. In this world, the power of digital information to catalyze progress is limited only by the power of the human mind. Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity. A few bits, well found, can drive a giant leap of creativity. The power of a data set is amplified by ingenuity through applications unimagined by the authors and distant from the original field.

Key characteristics of the current digital data landscape are:

- *the products of science and the starting point for new research are increasingly digital and increasingly "born-digital";*
- *exploding volumes and rising demand for data use are driven by the rapid pace of digital technology innovations;*
- *all sectors of society are stakeholders in digital preservation and access; and*
- *a comprehensive framework for cooperation and coordination to manage the risks to preservation of digital data is missing.*

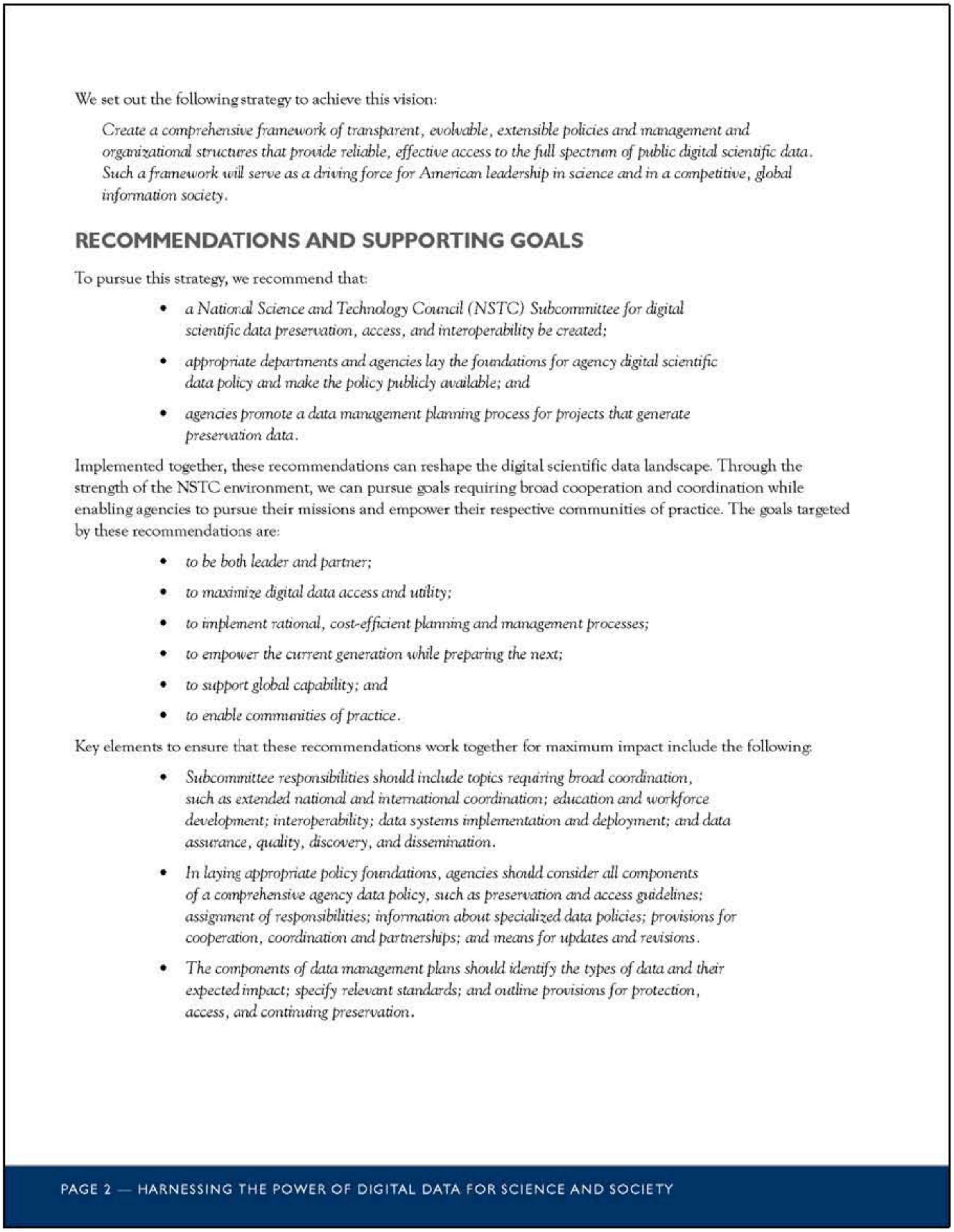
The following guiding principles were deduced from an analysis of the current digital scientific data landscape. These are based on the expertise of the members of the Interagency Working Group on Digital Data (IWGDD), supplemented by input from outside experts and documentation from major studies of the challenges and opportunities presented by a fully digital world. These guiding principles are:

- *science is global and thrives in the digital dimensions;*
- *digital scientific data are national and global assets;*
- *not all digital scientific data need to be preserved and not all preserved data need to be preserved indefinitely;*
- *communities of practice are an essential feature of the digital landscape;*
- *preservation of digital scientific data is both a government and private sector responsibility and benefits society as a whole;*
- *long-term preservation, access, and interoperability require management of the full data life cycle; and*
- *dynamic strategies are required.*

The strategic framework, recommendations, and goals presented in this report are founded on these guiding principles.

VISION AND STRATEGY

We envision a digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination can be appropriately, reliably, and readily managed. This will enhance the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society.



We set out the following strategy to achieve this vision:

> *Create a comprehensive framework of transparent, evolvable, extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data. Such a framework will serve as a driving force for American leadership in science and in a competitive, global information society.*

RECOMMENDATIONS AND SUPPORTING GOALS

To pursue this strategy, we recommend that:

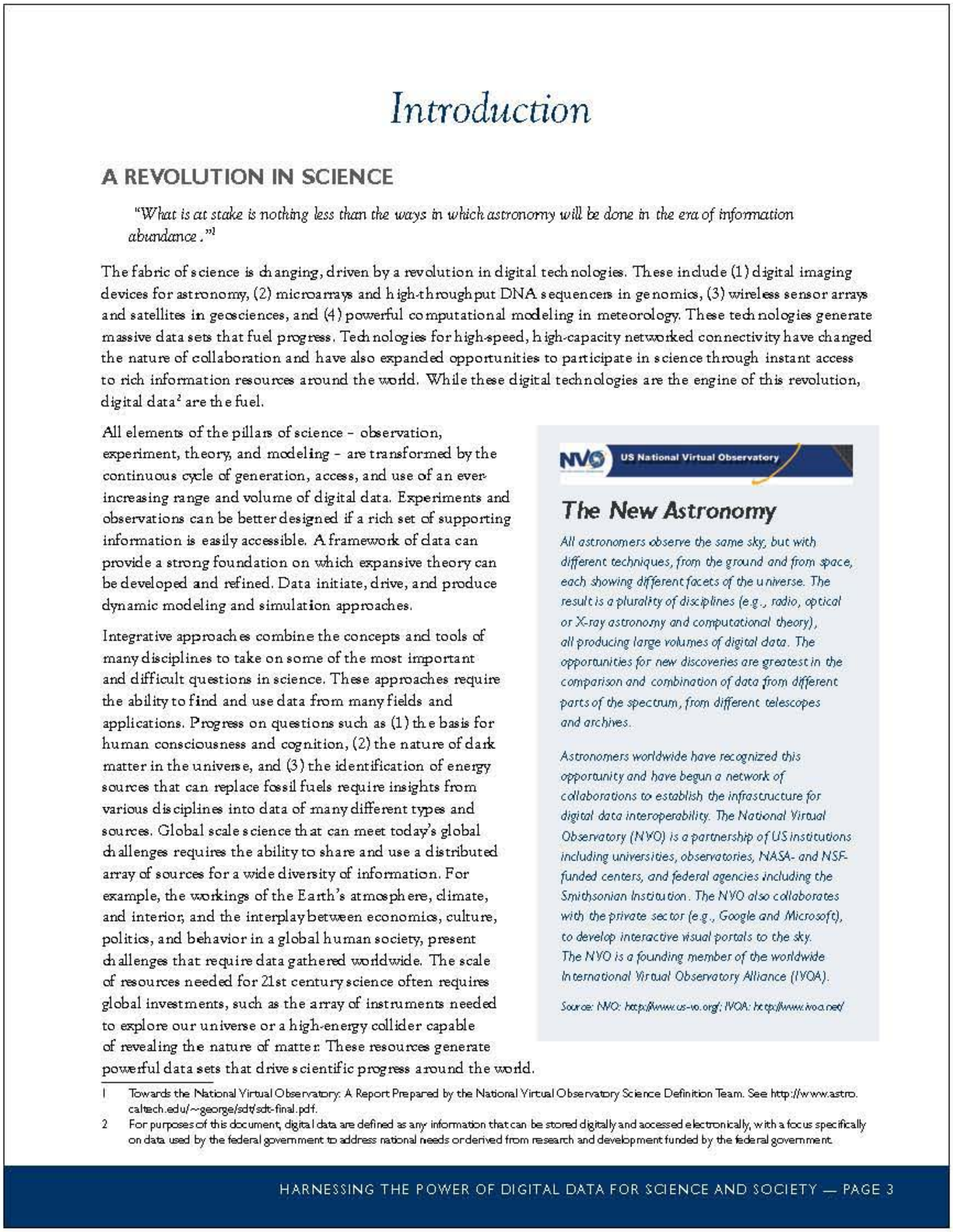
- *a National Science and Technology Council (NSTC) Subcommittee for digital scientific data preservation, access, and interoperability be created;*
- *appropriate departments and agencies lay the foundations for agency digital scientific data policy and make the policy publicly available; and*
- *agencies promote a data management planning process for projects that generate preservation data.*

Implemented together, these recommendations can reshape the digital scientific data landscape. Through the strength of the NSTC environment, we can pursue goals requiring broad cooperation and coordination while enabling agencies to pursue their missions and empower their respective communities of practice. The goals targeted by these recommendations are:

- *to be both leader and partner;*
- *to maximize digital data access and utility;*
- *to implement rational, cost-efficient planning and management processes;*
- *to empower the current generation while preparing the next;*
- *to support global capability; and*
- *to enable communities of practice.*

Key elements to ensure that these recommendations work together for maximum impact include the following:

- *Subcommittee responsibilities should include topics requiring broad coordination, such as extended national and international coordination; education and workforce development; interoperability; data systems implementation and deployment; and data assurance, quality, discovery, and dissemination.*
- *In laying appropriate policy foundations, agencies should consider all components of a comprehensive agency data policy, such as preservation and access guidelines; assignment of responsibilities; information about specialized data policies; provisions for cooperation, coordination and partnerships; and means for updates and revisions.*
- *The components of data management plans should identify the types of data and their expected impact; specify relevant standards; and outline provisions for protection, access, and continuing preservation.*



Introduction

A REVOLUTION IN SCIENCE

> *"What is at stake is nothing less than the ways in which astronomy will be done in the era of information abundance."*[1]

The fabric of science is changing, driven by a revolution in digital technologies. These include (1) digital imaging devices for astronomy, (2) microarrays and high-throughput DNA sequencers in genomics, (3) wireless sensor arrays and satellites in geosciences, and (4) powerful computational modeling in meteorology. These technologies generate massive data sets that fuel progress. Technologies for high-speed, high-capacity networked connectivity have changed the nature of collaboration and have also expanded opportunities to participate in science through instant access to rich information resources around the world. While these digital technologies are the engine of this revolution, digital data[2] are the fuel.

All elements of the pillars of science – observation, experiment, theory, and modeling – are transformed by the continuous cycle of generation, access, and use of an ever-increasing range and volume of digital data. Experiments and observations can be better designed if a rich set of supporting information is easily accessible. A framework of data can provide a strong foundation on which expansive theory can be developed and refined. Data initiate, drive, and produce dynamic modeling and simulation approaches.

Integrative approaches combine the concepts and tools of many disciplines to take on some of the most important and difficult questions in science. These approaches require the ability to find and use data from many fields and applications. Progress on questions such as (1) the basis for human consciousness and cognition, (2) the nature of dark matter in the universe, and (3) the identification of energy sources that can replace fossil fuels require insights from various disciplines into data of many different types and sources. Global scale science that can meet today's global challenges requires the ability to share and use a distributed array of sources for a wide diversity of information. For example, the workings of the Earth's atmosphere, climate, and interior, and the interplay between economics, culture, politics, and behavior in a global human society, present challenges that require data gathered worldwide. The scale of resources needed for 21st century science often requires global investments, such as the array of instruments needed to explore our universe or a high-energy collider capable of revealing the nature of matter. These resources generate powerful data sets that drive scientific progress around the world.

US National Virtual Observatory

The New Astronomy

All astronomers observe the same sky, but with different techniques, from the ground and from space, each showing different facets of the universe. The result is a plurality of disciplines (e.g., radio, optical or X-ray astronomy and computational theory), all producing large volumes of digital data. The opportunities for new discoveries are greatest in the comparison and combination of data from different parts of the spectrum, from different telescopes and archives.

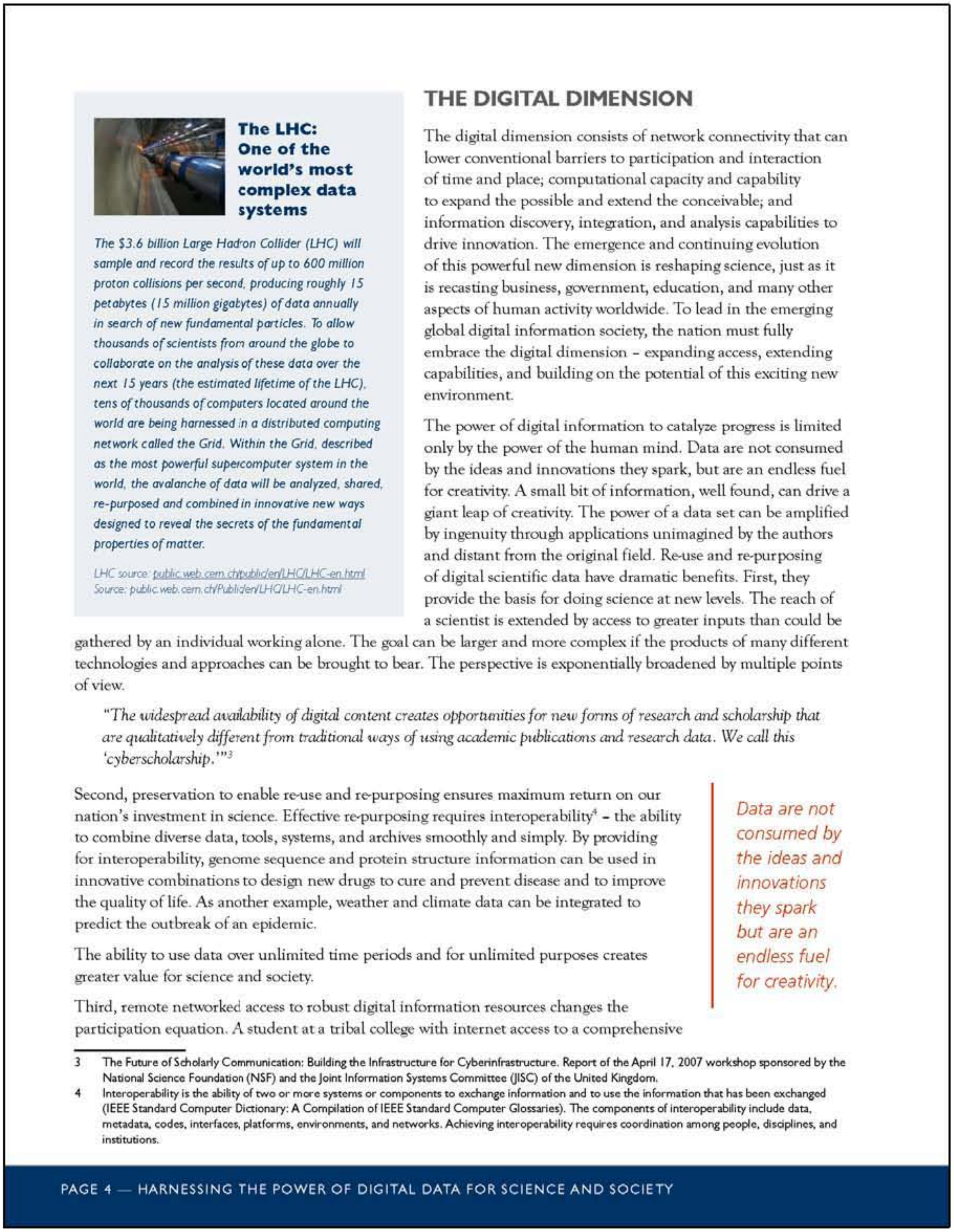
Astronomers worldwide have recognized this opportunity and have begun a network of collaborations to establish the infrastructure for digital data interoperability. The National Virtual Observatory (NVO) is a partnership of US institutions including universities, observatories, NASA- and NSF-funded centers, and federal agencies including the Smithsonian Institution. The NVO also collaborates with the private sector (e.g., Google and Microsoft), to develop interactive visual portals to the sky. The NVO is a founding member of the worldwide International Virtual Observatory Alliance (IVOA).

Source: NVO: http://www.us-vo.org/; IVOA: http://www.ivoa.net/

[1] Towards the National Virtual Observatory: A Report Prepared by the National Virtual Observatory Science Definition Team. See http://www.astro.caltech.edu/~george/sdt/sdt-final.pdf.

[2] For purposes of this document, digital data are defined as any information that can be stored digitally and accessed electronically, with a focus specifically on data used by the federal government to address national needs or derived from research and development funded by the federal government.

HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY — PAGE 3



The LHC: One of the world's most complex data systems

The $3.6 billion Large Hadron Collider (LHC) will sample and record the results of up to 600 million proton collisions per second, producing roughly 15 petabytes (15 million gigabytes) of data annually in search of new fundamental particles. To allow thousands of scientists from around the globe to collaborate on the analysis of these data over the next 15 years (the estimated lifetime of the LHC), tens of thousands of computers located around the world are being harnessed in a distributed computing network called the Grid. Within the Grid, described as the most powerful supercomputer system in the world, the avalanche of data will be analyzed, shared, re-purposed and combined in innovative new ways designed to reveal the secrets of the fundamental properties of matter.

LHC source: public.web.cern.ch/public/en/LHC/LHC-en.html
Source: public.web.cern.ch/Public/en/LHC/LHC-en.html

THE DIGITAL DIMENSION

The digital dimension consists of network connectivity that can lower conventional barriers to participation and interaction of time and place; computational capacity and capability to expand the possible and extend the conceivable; and information discovery, integration, and analysis capabilities to drive innovation. The emergence and continuing evolution of this powerful new dimension is reshaping science, just as it is recasting business, government, education, and many other aspects of human activity worldwide. To lead in the emerging global digital information society, the nation must fully embrace the digital dimension – expanding access, extending capabilities, and building on the potential of this exciting new environment.

The power of digital information to catalyze progress is limited only by the power of the human mind. Data are not consumed by the ideas and innovations they spark, but are an endless fuel for creativity. A small bit of information, well found, can drive a giant leap of creativity. The power of a data set can be amplified by ingenuity through applications unimagined by the authors and distant from the original field. Re-use and re-purposing of digital scientific data have dramatic benefits. First, they provide the basis for doing science at new levels. The reach of a scientist is extended by access to greater inputs than could be gathered by an individual working alone. The goal can be larger and more complex if the products of many different technologies and approaches can be brought to bear. The perspective is exponentially broadened by multiple points of view.

> *"The widespread availability of digital content creates opportunities for new forms of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data. We call this 'cyberscholarship.'"*[3]

Second, preservation to enable re-use and re-purposing ensures maximum return on our nation's investment in science. Effective re-purposing requires interoperability[4] – the ability to combine diverse data, tools, systems, and archives smoothly and simply. By providing for interoperability, genome sequence and protein structure information can be used in innovative combinations to design new drugs to cure and prevent disease and to improve the quality of life. As another example, weather and climate data can be integrated to predict the outbreak of an epidemic.

Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity.

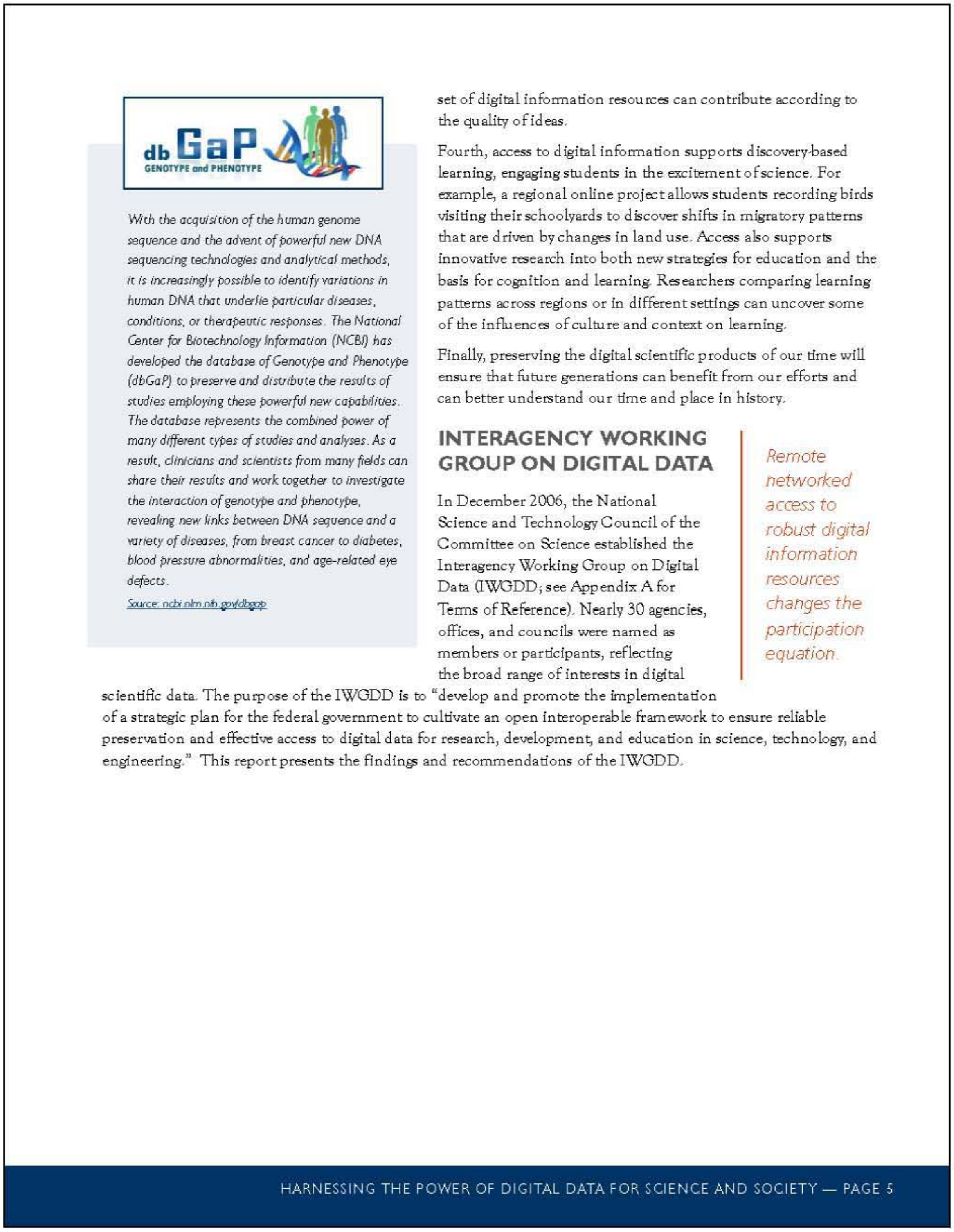
The ability to use data over unlimited time periods and for unlimited purposes creates greater value for science and society.

Third, remote networked access to robust digital information resources changes the participation equation. A student at a tribal college with internet access to a comprehensive

[3] The Future of Scholarly Communication: Building the Infrastructure for Cyberinfrastructure. Report of the April 17, 2007 workshop sponsored by the National Science Foundation (NSF) and the Joint Information Systems Committee (JISC) of the United Kingdom.

[4] Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged (IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries). The components of interoperability include data, metadata, codes, interfaces, platforms, environments, and networks. Achieving interoperability requires coordination among people, disciplines, and institutions.

PAGE 4 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



With the acquisition of the human genome sequence and the advent of powerful new DNA sequencing technologies and analytical methods, it is increasingly possible to identify variations in human DNA that underlie particular diseases, conditions, or therapeutic responses. The National Center for Biotechnology Information (NCBI) has developed the database of Genotype and Phenotype (dbGaP) to preserve and distribute the results of studies employing these powerful new capabilities. The database represents the combined power of many different types of studies and analyses. As a result, clinicians and scientists from many fields can share their results and work together to investigate the interaction of genotype and phenotype, revealing new links between DNA sequence and a variety of diseases, from breast cancer to diabetes, blood pressure abnormalities, and age-related eye defects.

Source: ncbi.nlm.nih.gov/dbgap

set of digital information resources can contribute according to the quality of ideas.

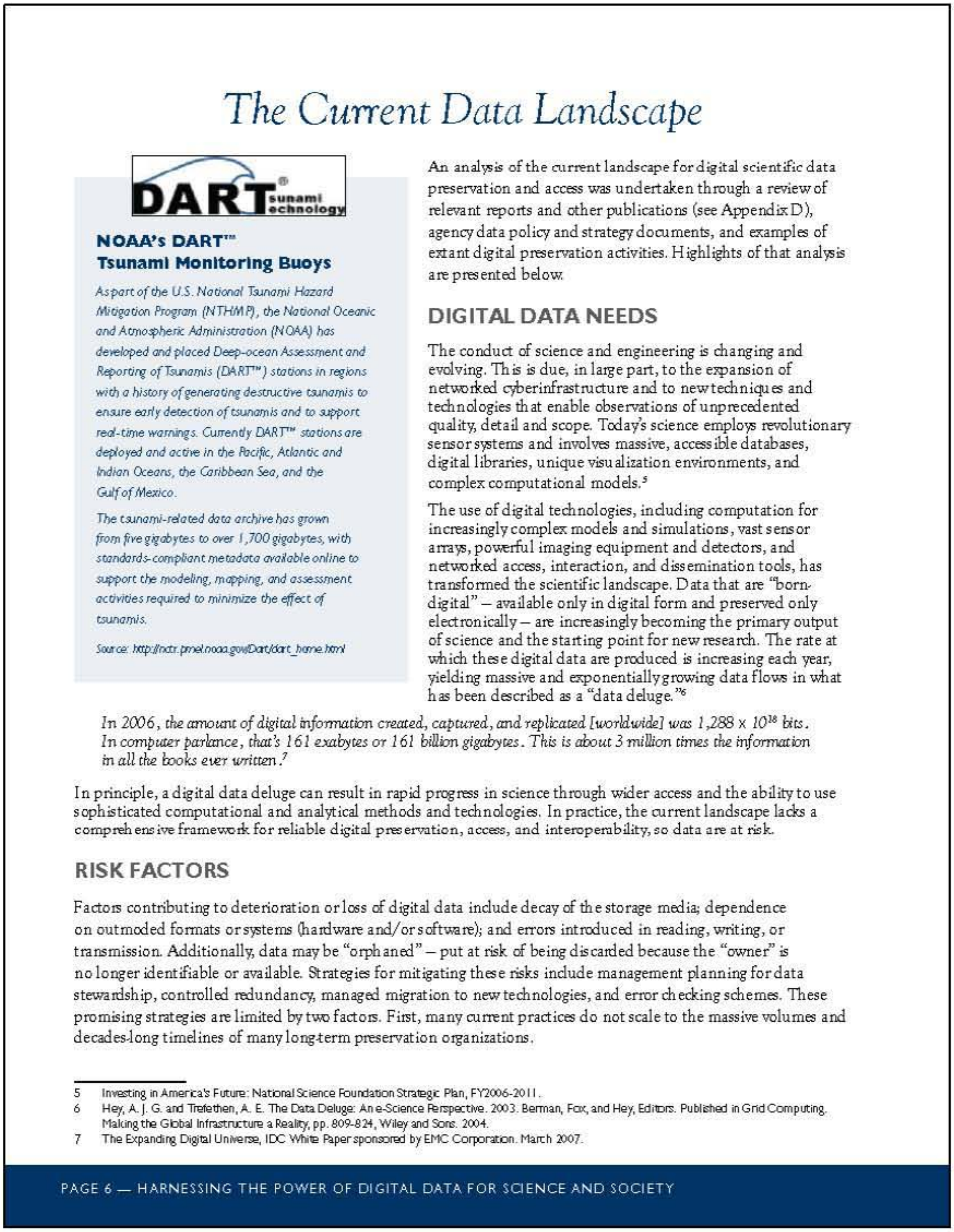
Fourth, access to digital information supports discovery-based learning, engaging students in the excitement of science. For example, a regional online project allows students recording birds visiting their schoolyards to discover shifts in migratory patterns that are driven by changes in land use. Access also supports innovative research into both new strategies for education and the basis for cognition and learning. Researchers comparing learning patterns across regions or in different settings can uncover some of the influences of culture and context on learning.

Finally, preserving the digital scientific products of our time will ensure that future generations can benefit from our efforts and can better understand our time and place in history.

INTERAGENCY WORKING GROUP ON DIGITAL DATA

In December 2006, the National Science and Technology Council of the Committee on Science established the Interagency Working Group on Digital Data (IWGDD; see Appendix A for Terms of Reference). Nearly 30 agencies, offices, and councils were named as members or participants, reflecting the broad range of interests in digital scientific data. The purpose of the IWGDD is to "develop and promote the implementation of a strategic plan for the federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering." This report presents the findings and recommendations of the IWGDD.

Remote networked access to robust digital information resources changes the participation equation.



The Current Data Landscape

NOAA's DART™ Tsunami Monitoring Buoys

As part of the U.S. National Tsunami Hazard Mitigation Program (NTHMP), the National Oceanic and Atmospheric Administration (NOAA) has developed and placed Deep-ocean Assessment and Reporting of Tsunamis (DART™) stations in regions with a history of generating destructive tsunamis to ensure early detection of tsunamis and to support real-time warnings. Currently DART™ stations are deployed and active in the Pacific, Atlantic and Indian Oceans, the Caribbean Sea, and the Gulf of Mexico.

The tsunami-related data archive has grown from five gigabytes to over 1,700 gigabytes, with standards-compliant metadata available online to support the modeling, mapping, and assessment activities required to minimize the effect of tsunamis.

Source: http://nctr.pmel.noaa.gov/Dart/dart_home.html

An analysis of the current landscape for digital scientific data preservation and access was undertaken through a review of relevant reports and other publications (see Appendix D), agency data policy and strategy documents, and examples of extant digital preservation activities. Highlights of that analysis are presented below.

DIGITAL DATA NEEDS

The conduct of science and engineering is changing and evolving. This is due, in large part, to the expansion of networked cyberinfrastructure and to new techniques and technologies that enable observations of unprecedented quality, detail and scope. Today's science employs revolutionary sensor systems and involves massive, accessible databases, digital libraries, unique visualization environments, and complex computational models.[5]

The use of digital technologies, including computation for increasingly complex models and simulations, vast sensor arrays, powerful imaging equipment and detectors, and networked access, interaction, and dissemination tools, has transformed the scientific landscape. Data that are "born-digital" – available only in digital form and preserved only electronically – are increasingly becoming the primary output of science and the starting point for new research. The rate at which these digital data are produced is increasing each year, yielding massive and exponentially growing data flows in what has been described as a "data deluge."[6]

> *In 2006, the amount of digital information created, captured, and replicated [worldwide] was 1,288 x 10^{16} bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes. This is about 3 million times the information in all the books ever written.*[7]

In principle, a digital data deluge can result in rapid progress in science through wider access and the ability to use sophisticated computational and analytical methods and technologies. In practice, the current landscape lacks a comprehensive framework for reliable digital preservation, access, and interoperability, so data are at risk.

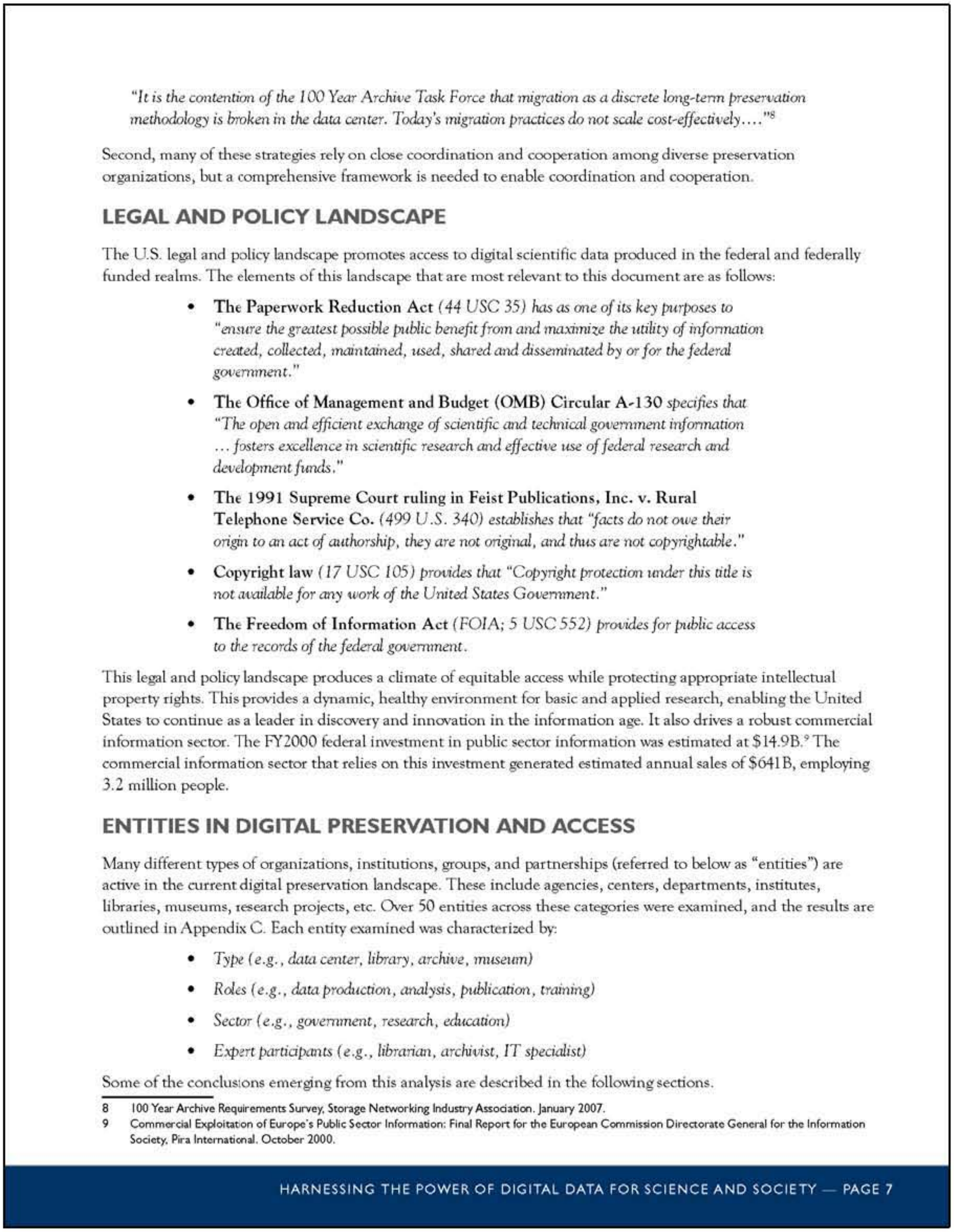
RISK FACTORS

Factors contributing to deterioration or loss of digital data include decay of the storage media; dependence on outmoded formats or systems (hardware and/or software); and errors introduced in reading, writing, or transmission. Additionally, data may be "orphaned" – put at risk of being discarded because the "owner" is no longer identifiable or available. Strategies for mitigating these risks include management planning for data stewardship, controlled redundancy, managed migration to new technologies, and error checking schemes. These promising strategies are limited by two factors. First, many current practices do not scale to the massive volumes and decades-long timelines of many long-term preservation organizations.

[5] Investing in America's Future: National Science Foundation Strategic Plan, FY2006-2011.

[6] Hey, A. J. G. and Trefethen, A. E. The Data Deluge: An e-Science Perspective. 2003. Berman, Fox, and Hey, Editors. Published in Grid Computing. Making the Global Infrastructure a Reality, pp. 809-824, Wiley and Sons. 2004.

[7] The Expanding Digital Universe, IDC White Paper sponsored by EMC Corporation. March 2007.



> *"It is the contention of the 100 Year Archive Task Force that migration as a discrete long-term preservation methodology is broken in the data center. Today's migration practices do not scale cost-effectively…."*[8]

Second, many of these strategies rely on close coordination and cooperation among diverse preservation organizations, but a comprehensive framework is needed to enable coordination and cooperation.

LEGAL AND POLICY LANDSCAPE

The U.S. legal and policy landscape promotes access to digital scientific data produced in the federal and federally funded realms. The elements of this landscape that are most relevant to this document are as follows:

- **The Paperwork Reduction Act** *(44 USC 35) has as one of its key purposes to "ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the federal government."*
- **The Office of Management and Budget (OMB) Circular A-130** *specifies that "The open and efficient exchange of scientific and technical government information … fosters excellence in scientific research and effective use of federal research and development funds."*
- **The 1991 Supreme Court ruling in Feist Publications, Inc. v. Rural Telephone Service Co.** *(499 U.S. 340) establishes that "facts do not owe their origin to an act of authorship, they are not original, and thus are not copyrightable."*
- **Copyright law** *(17 USC 105) provides that "Copyright protection under this title is not available for any work of the United States Government."*
- **The Freedom of Information Act** *(FOIA; 5 USC 552) provides for public access to the records of the federal government.*

This legal and policy landscape produces a climate of equitable access while protecting appropriate intellectual property rights. This provides a dynamic, healthy environment for basic and applied research, enabling the United States to continue as a leader in discovery and innovation in the information age. It also drives a robust commercial information sector. The FY2000 federal investment in public sector information was estimated at $14.9B.[9] The commercial information sector that relies on this investment generated estimated annual sales of $641B, employing 3.2 million people.

ENTITIES IN DIGITAL PRESERVATION AND ACCESS

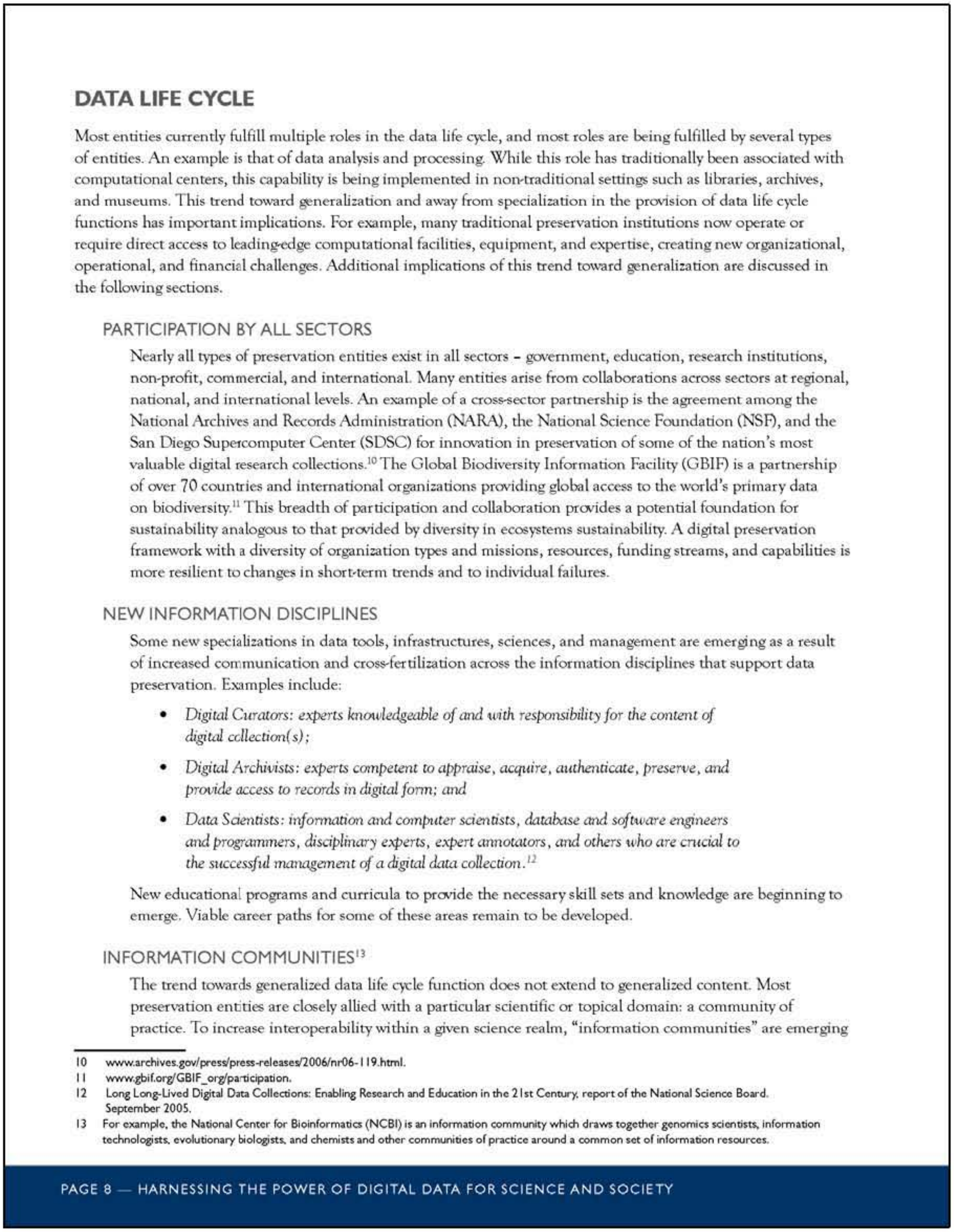
Many different types of organizations, institutions, groups, and partnerships (referred to below as "entities") are active in the current digital preservation landscape. These include agencies, centers, departments, institutes, libraries, museums, research projects, etc. Over 50 entities across these categories were examined, and the results are outlined in Appendix C. Each entity examined was characterized by:

- *Type (e.g., data center, library, archive, museum)*
- *Roles (e.g., data production, analysis, publication, training)*
- *Sector (e.g., government, research, education)*
- *Expert participants (e.g., librarian, archivist, IT specialist)*

Some of the conclusions emerging from this analysis are described in the following sections.

[8] 100 Year Archive Requirements Survey, Storage Networking Industry Association. January 2007.

[9] Commercial Exploitation of Europe's Public Sector Information: Final Report for the European Commission Directorate General for the Information Society, Pira International. October 2000.



DATA LIFE CYCLE

Most entities currently fulfill multiple roles in the data life cycle, and most roles are being fulfilled by several types of entities. An example is that of data analysis and processing. While this role has traditionally been associated with computational centers, this capability is being implemented in non-traditional settings such as libraries, archives, and museums. This trend toward generalization and away from specialization in the provision of data life cycle functions has important implications. For example, many traditional preservation institutions now operate or require direct access to leading-edge computational facilities, equipment, and expertise, creating new organizational, operational, and financial challenges. Additional implications of this trend toward generalization are discussed in the following sections.

PARTICIPATION BY ALL SECTORS

Nearly all types of preservation entities exist in all sectors – government, education, research institutions, non-profit, commercial, and international. Many entities arise from collaborations across sectors at regional, national, and international levels. An example of a cross-sector partnership is the agreement among the National Archives and Records Administration (NARA), the National Science Foundation (NSF), and the San Diego Supercomputer Center (SDSC) for innovation in preservation of some of the nation's most valuable digital research collections.[10] The Global Biodiversity Information Facility (GBIF) is a partnership of over 70 countries and international organizations providing global access to the world's primary data on biodiversity.[11] This breadth of participation and collaboration provides a potential foundation for sustainability analogous to that provided by diversity in ecosystems sustainability. A digital preservation framework with a diversity of organization types and missions, resources, funding streams, and capabilities is more resilient to changes in short-term trends and to individual failures.

NEW INFORMATION DISCIPLINES

Some new specializations in data tools, infrastructures, sciences, and management are emerging as a result of increased communication and cross-fertilization across the information disciplines that support data preservation. Examples include:

- *Digital Curators: experts knowledgeable of and with responsibility for the content of digital collection(s);*
- *Digital Archivists: experts competent to appraise, acquire, authenticate, preserve, and provide access to records in digital form; and*
- *Data Scientists: information and computer scientists, database and software engineers and programmers, disciplinary experts, expert annotators, and others who are crucial to the successful management of a digital data collection.*[12]

New educational programs and curricula to provide the necessary skill sets and knowledge are beginning to emerge. Viable career paths for some of these areas remain to be developed.

INFORMATION COMMUNITIES[13]

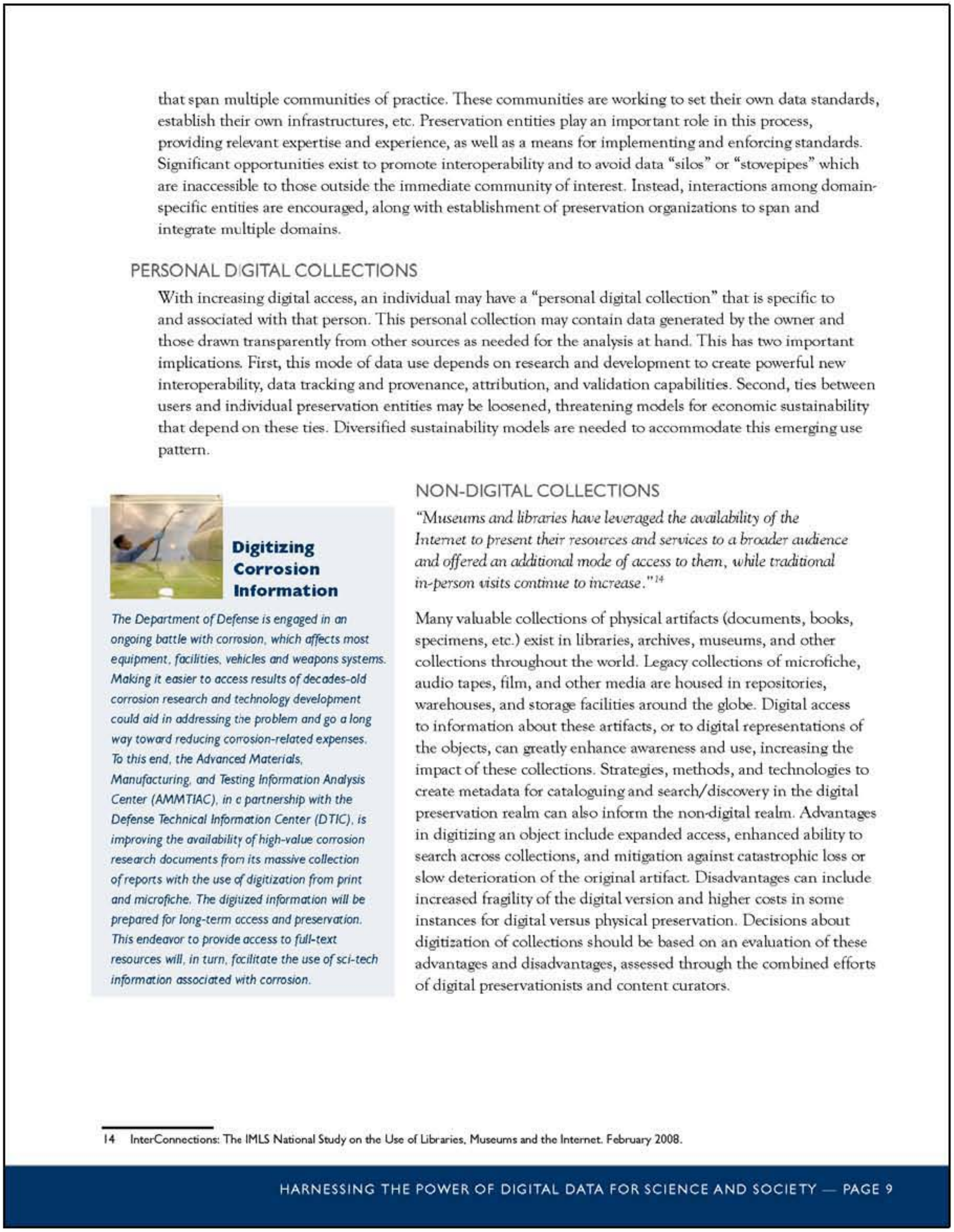
The trend towards generalized data life cycle function does not extend to generalized content. Most preservation entities are closely allied with a particular scientific or topical domain: a community of practice. To increase interoperability within a given science realm, "information communities" are emerging

[10] www.archives.gov/press/press-releases/2006/nr06-119.html.

[11] www.gbif.org/GBIF_org/participation.

[12] Long Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, report of the National Science Board. September 2005.

[13] For example, the National Center for Bioinformatics (NCBI) is an information community which draws together genomics scientists, information technologists, evolutionary biologists, and chemists and other communities of practice around a common set of information resources.



that span multiple communities of practice. These communities are working to set their own data standards, establish their own infrastructures, etc. Preservation entities play an important role in this process, providing relevant expertise and experience, as well as a means for implementing and enforcing standards. Significant opportunities exist to promote interoperability and to avoid data "silos" or "stovepipes" which are inaccessible to those outside the immediate community of interest. Instead, interactions among domain-specific entities are encouraged, along with establishment of preservation organizations to span and integrate multiple domains.

PERSONAL DIGITAL COLLECTIONS

With increasing digital access, an individual may have a "personal digital collection" that is specific to and associated with that person. This personal collection may contain data generated by the owner and those drawn transparently from other sources as needed for the analysis at hand. This has two important implications. First, this mode of data use depends on research and development to create powerful new interoperability, data tracking and provenance, attribution, and validation capabilities. Second, ties between users and individual preservation entities may be loosened, threatening models for economic sustainability that depend on these ties. Diversified sustainability models are needed to accommodate this emerging use pattern.

NON-DIGITAL COLLECTIONS

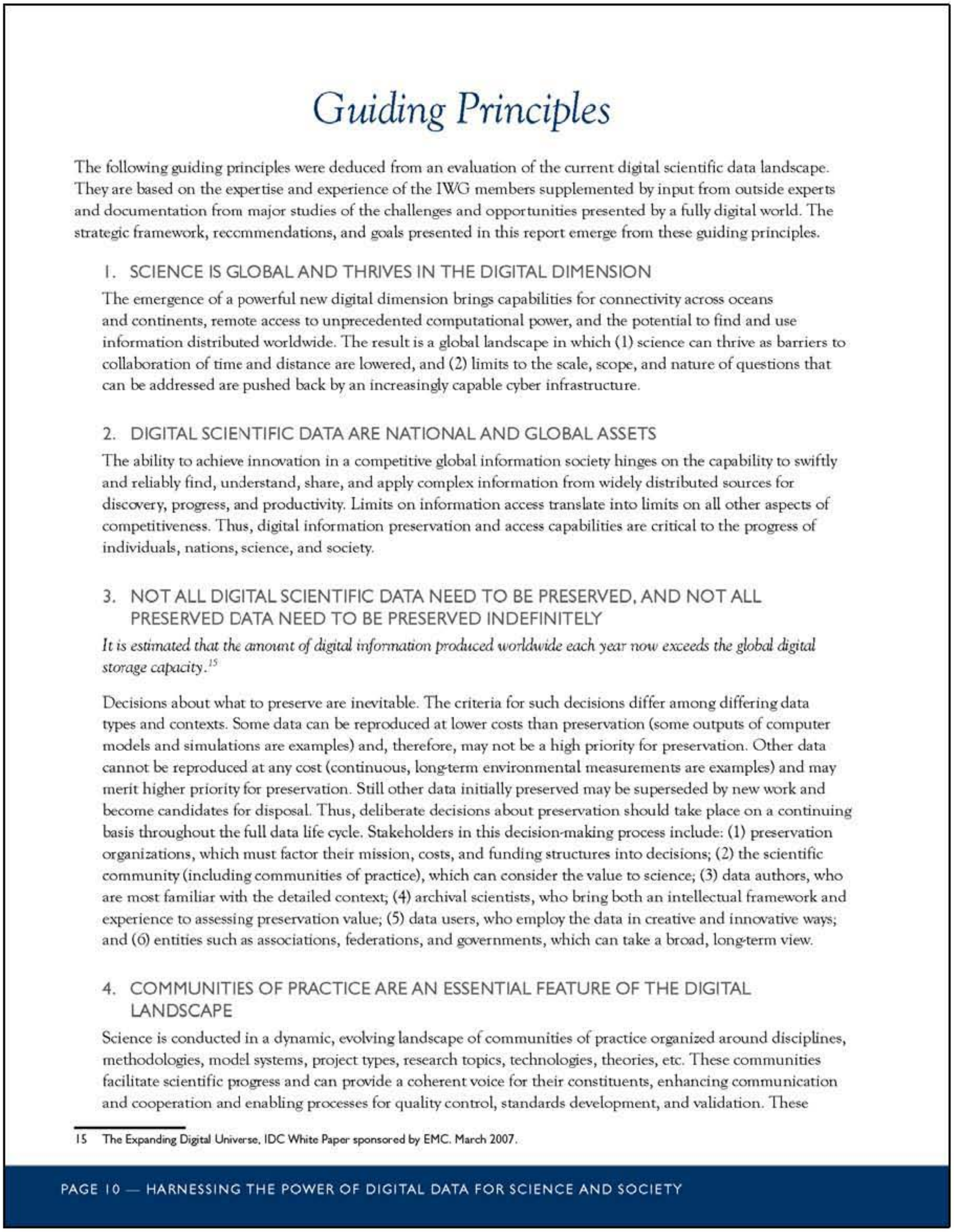
> *"Museums and libraries have leveraged the availability of the Internet to present their resources and services to a broader audience and offered an additional mode of access to them, while traditional in-person visits continue to increase."*[14]

Many valuable collections of physical artifacts (documents, books, specimens, etc.) exist in libraries, archives, museums, and other collections throughout the world. Legacy collections of microfiche, audio tapes, film, and other media are housed in repositories, warehouses, and storage facilities around the globe. Digital access to information about these artifacts, or to digital representations of the objects, can greatly enhance awareness and use, increasing the impact of these collections. Strategies, methods, and technologies to create metadata for cataloguing and search/discovery in the digital preservation realm can also inform the non-digital realm. Advantages in digitizing an object include expanded access, enhanced ability to search across collections, and mitigation against catastrophic loss or slow deterioration of the original artifact. Disadvantages can include increased fragility of the digital version and higher costs in some instances for digital versus physical preservation. Decisions about digitization of collections should be based on an evaluation of these advantages and disadvantages, assessed through the combined efforts of digital preservationists and content curators.

Digitizing Corrosion Information

The Department of Defense is engaged in an ongoing battle with corrosion, which affects most equipment, facilities, vehicles and weapons systems. Making it easier to access results of decades-old corrosion research and technology development could aid in addressing the problem and go a long way toward reducing corrosion-related expenses. To this end, the Advanced Materials, Manufacturing, and Testing Information Analysis Center (AMMTIAC), in a partnership with the Defense Technical Information Center (DTIC), is improving the availability of high-value corrosion research documents from its massive collection of reports with the use of digitization from print and microfiche. The digitized information will be prepared for long-term access and preservation. This endeavor to provide access to full-text resources will, in turn, facilitate the use of sci-tech information associated with corrosion.

[14] InterConnections: The IMLS National Study on the Use of Libraries, Museums and the Internet. February 2008.



Guiding Principles

The following guiding principles were deduced from an evaluation of the current digital scientific data landscape. They are based on the expertise and experience of the IWG members supplemented by input from outside experts and documentation from major studies of the challenges and opportunities presented by a fully digital world. The strategic framework, reccmmendations, and goals presented in this report emerge from these guiding principles.

1. SCIENCE IS GLOBAL AND THRIVES IN THE DIGITAL DIMENSION

The emergence of a powerful new digital dimension brings capabilities for connectivity across oceans and continents, remote access to unprecedented computational power, and the potential to find and use information distributed worldwide. The result is a global landscape in which (1) science can thrive as barriers to collaboration of time and distance are lowered, and (2) limits to the scale, scope, and nature of questions that can be addressed are pushed back by an increasingly capable cyber infrastructure.

2. DIGITAL SCIENTIFIC DATA ARE NATIONAL AND GLOBAL ASSETS

The ability to achieve innovation in a competitive global information society hinges on the capability to swiftly and reliably find, understand, share, and apply complex information from widely distributed sources for discovery, progress, and productivity. Limits on information access translate into limits on all other aspects of competitiveness. Thus, digital information preservation and access capabilities are critical to the progress of individuals, nations, science, and society.

3. NOT ALL DIGITAL SCIENTIFIC DATA NEED TO BE PRESERVED, AND NOT ALL PRESERVED DATA NEED TO BE PRESERVED INDEFINITELY

It is estimated that the amount of digital information produced worldwide each year now exceeds the global digital storage capacity.[15]

Decisions about what to preserve are inevitable. The criteria for such decisions differ among differing data types and contexts. Some data can be reproduced at lower costs than preservation (some outputs of computer models and simulations are examples) and, therefore, may not be a high priority for preservation. Other data cannot be reproduced at any cost (continuous, long-term environmental measurements are examples) and may merit higher priority for preservation. Still other data initially preserved may be superseded by new work and become candidates for disposal. Thus, deliberate decisions about preservation should take place on a continuing basis throughout the full data life cycle. Stakeholders in this decision-making process include: (1) preservation organizations, which must factor their mission, costs, and funding structures into decisions; (2) the scientific community (including communities of practice), which can consider the value to science; (3) data authors, who are most familiar with the detailed context; (4) archival scientists, who bring both an intellectual framework and experience to assessing preservation value; (5) data users, who employ the data in creative and innovative ways; and (6) entities such as associations, federations, and governments, which can take a broad, long-term view.

4. COMMUNITIES OF PRACTICE ARE AN ESSENTIAL FEATURE OF THE DIGITAL LANDSCAPE

Science is conducted in a dynamic, evolving landscape of communities of practice organized around disciplines, methodologies, model systems, project types, research topics, technologies, theories, etc. These communities facilitate scientific progress and can provide a coherent voice for their constituents, enhancing communication and cooperation and enabling processes for quality control, standards development, and validation. These

[15] The Expanding Digital Universe, IDC White Paper sponsored by EMC. March 2007.

capabilities are crucial for data preservation and access in communicating the needs and expectations of a community of users, providing expert input on the scientific context for data (including input to decisions about what to preserve and what to discard), promoting good data management practices, and contributing to the development of effective data standards. Thus, data preservation policies and strategies must encourage and enable communities of practice both because of their important role in science and because of the capabilities and perspectives they bring to the preservation process. "One-size-fits-all" policies must be avoided to allow for strategies and designing mechanisms for interoperability that support communities of practice.

Barcode of Life

The Barcode of Life Initiative is an international effort to develop reliable and authoritative means for the global identification of biological species. Barcoding uses a short DNA sequence within an organism's genome as the equivalent of a barcode on a supermarket product to determine the species origin of a biological sample. Adoption of a standard format for barcode data allows a sample in a museum or collected in the field to be instantly linked to related information resources worldwide; to be tied in to relevant tissue, parasite, and other collections globally; and to reference DNA databases in the United States, Japan, and Europe. The result is the ability to conduct biodiversity, species migration and invasion, and population genetics studies that are more powerful because they can be reliably compared to and informed by other projects worldwide.

Source: http://barcoding.si.edu/

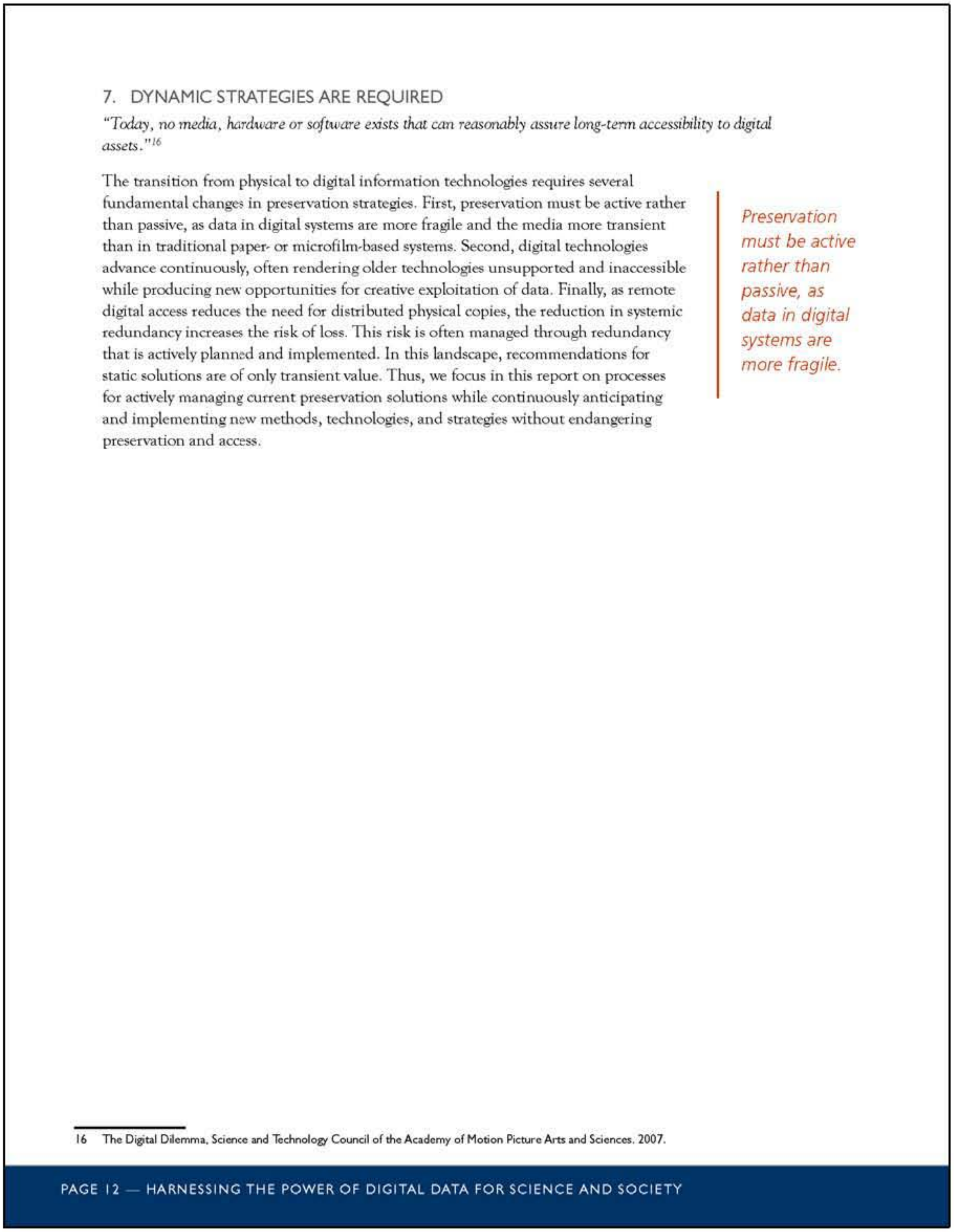
5. PRESERVATION OF DIGITAL SCIENTIFIC DATA IS BOTH A GOVERNMENT AND PRIVATE SECTOR RESPONSIBILITY, AND BENEFITS SOCIETY AS A WHOLE

A large number and wide variety of entities, organizations, and communities – each with their own assumptions, culture, expertise, objectives, policies, and resources – are involved in the creation and preservation for access of scientific digital data. Responsibilities for data stewardship are distributed across many diverse entities that, in turn, engage with different institutions, disciplines, and interdisciplinary domains. Responsibility for data stewardship should remain with the distributed collections and repositories that have a vested interest in their community's data. A framework of government/private sector partnerships (analogous to the air transportation or monetary systems) is required to link these distributed responsibilities into an effective system for digital preservation and access.

6. LONG-TERM PRESERVATION, ACCESS, AND INTEROPERABILITY REQUIRE MANAGEMENT OF THE FULL DATA LIFE CYCLE

The full data life cycle includes creation, ingestion or acquisition, documentation, organization, migration, protection, access, and disposition (see Appendix B for a description of the data life cycle) and has two important features. First, the cycle is dynamic rather than static and includes ongoing processes of curation, disposition, and use. Many processes, such as data analyses involving transformation or recombination, are catalytic, continuously increasing the volume of data for preservation and access. Second, the steps in the cycle are not independent. Feasibility, costs, and limitations for each step are strongly dependent on actions taken at other steps. For example, inadequate documentation at an early stage can prevent later use; failure to migrate to new technologies can leave data inaccessible. Effective management of each step and coordination across steps in the life cycle are required to ensure that data are reliably preserved and can be accessed and used efficiently.

Responsibilities for data stewardship are distributed across many diverse entities.



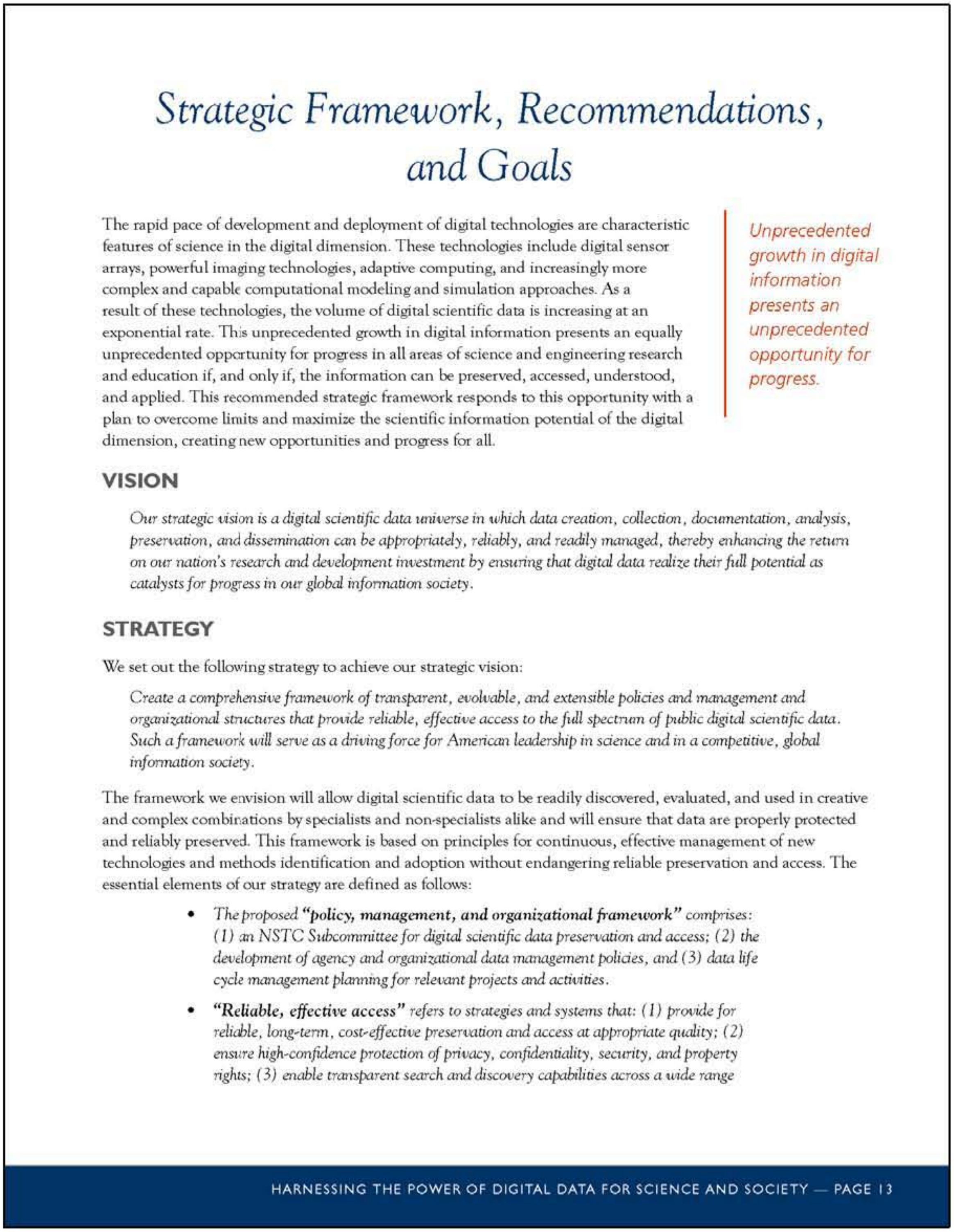
7. DYNAMIC STRATEGIES ARE REQUIRED

"Today, no media, hardware or software exists that can reasonably assure long-term accessibility to digital assets."[16]

The transition from physical to digital information technologies requires several fundamental changes in preservation strategies. First, preservation must be active rather than passive, as data in digital systems are more fragile and the media more transient than in traditional paper- or microfilm-based systems. Second, digital technologies advance continuously, often rendering older technologies unsupported and inaccessible while producing new opportunities for creative exploitation of data. Finally, as remote digital access reduces the need for distributed physical copies, the reduction in systemic redundancy increases the risk of loss. This risk is often managed through redundancy that is actively planned and implemented. In this landscape, recommendations for static solutions are of only transient value. Thus, we focus in this report on processes for actively managing current preservation solutions while continuously anticipating and implementing new methods, technologies, and strategies without endangering preservation and access.

Preservation must be active rather than passive, as data in digital systems are more fragile.

[16] The Digital Dilemma, Science and Technology Council of the Academy of Motion Picture Arts and Sciences. 2007.



Strategic Framework, Recommendations, and Goals

Unprecedented growth in digital information presents an unprecedented opportunity for progress.

The rapid pace of development and deployment of digital technologies are characteristic features of science in the digital dimension. These technologies include digital sensor arrays, powerful imaging technologies, adaptive computing, and increasingly more complex and capable computational modeling and simulation approaches. As a result of these technologies, the volume of digital scientific data is increasing at an exponential rate. This unprecedented growth in digital information presents an equally unprecedented opportunity for progress in all areas of science and engineering research and education if, and only if, the information can be preserved, accessed, understood, and applied. This recommended strategic framework responds to this opportunity with a plan to overcome limits and maximize the scientific information potential of the digital dimension, creating new opportunities and progress for all.

VISION

> *Our strategic vision is a digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination can be appropriately, reliably, and readily managed, thereby enhancing the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society.*

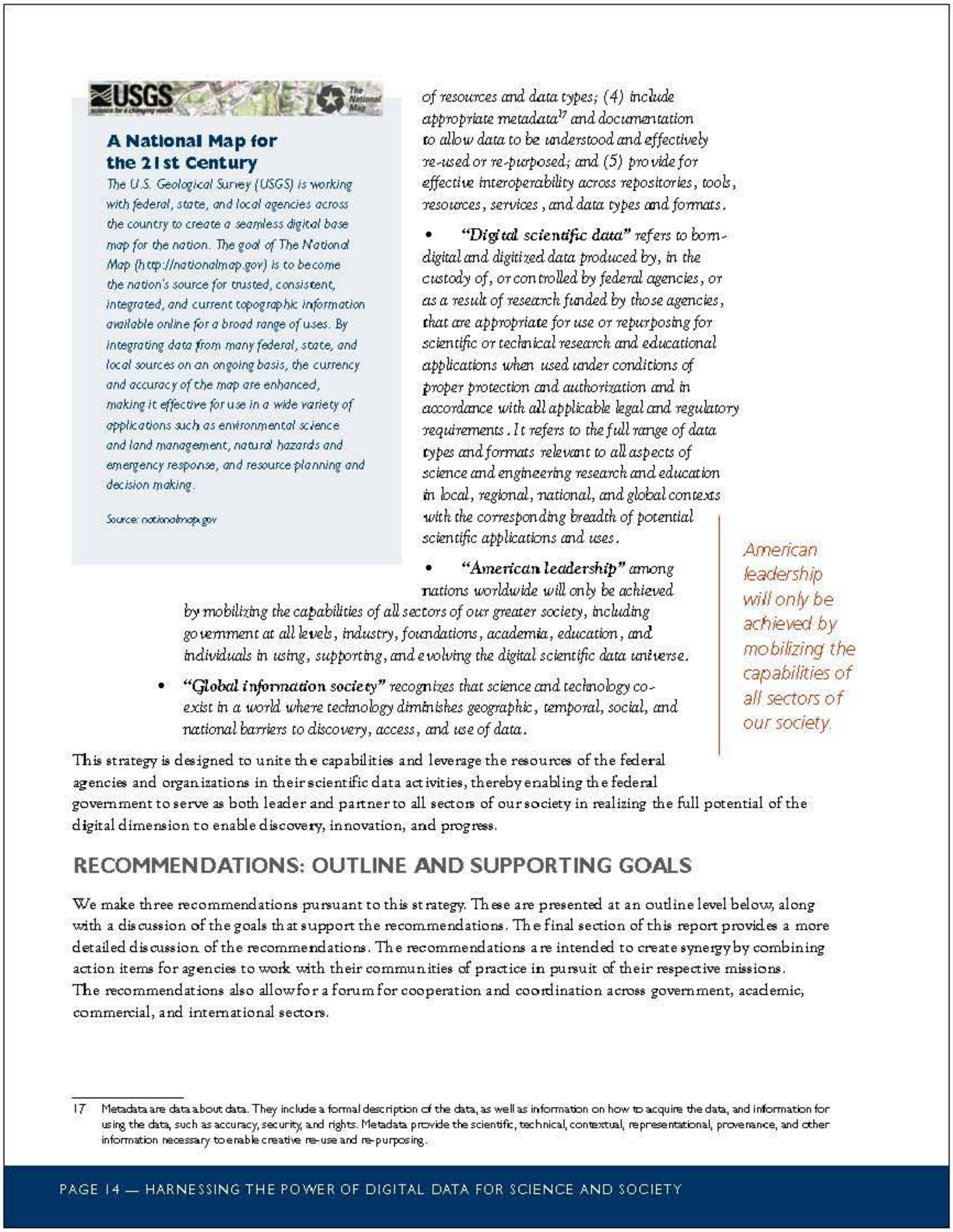
STRATEGY

We set out the following strategy to achieve our strategic vision:

> *Create a comprehensive framework of transparent, evolvable, and extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data. Such a framework will serve as a driving force for American leadership in science and in a competitive, global information society.*

The framework we envision will allow digital scientific data to be readily discovered, evaluated, and used in creative and complex combinations by specialists and non-specialists alike and will ensure that data are properly protected and reliably preserved. This framework is based on principles for continuous, effective management of new technologies and methods identification and adoption without endangering reliable preservation and access. The essential elements of our strategy are defined as follows:

- *The proposed* **"policy, management, and organizational framework"** *comprises: (1) an NSTC Subcommittee for digital scientific data preservation and access; (2) the development of agency and organizational data management policies, and (3) data life cycle management planning for relevant projects and activities.*
- **"Reliable, effective access"** *refers to strategies and systems that: (1) provide for reliable, long-term, cost-effective preservation and access at appropriate quality; (2) ensure high-confidence protection of privacy, confidentiality, security, and property rights; (3) enable transparent search and discovery capabilities across a wide range*



A National Map for the 21st Century

The U.S. Geological Survey (USGS) is working with federal, state, and local agencies across the country to create a seamless digital base map for the nation. The goal of The National Map (http://nationalmap.gov) is to become the nation's source for trusted, consistent, integrated, and current topographic information available online for a broad range of uses. By integrating data from many federal, state, and local sources on an ongoing basis, the currency and accuracy of the map are enhanced, making it effective for use in a wide variety of applications such as environmental science and land management, natural hazards and emergency response, and resource planning and decision making.

Source: nationalmap.gov

of resources and data types; (4) include appropriate metadata[17] and documentation to allow data to be understood and effectively re-used or re-purposed; and (5) provide for effective interoperability across repositories, tools, resources, services, and data types and formats.

- ***"Digital scientific data"*** *refers to born-digital and digitized data produced by, in the custody of, or controlled by federal agencies, or as a result of research funded by those agencies, that are appropriate for use or repurposing for scientific or technical research and educational applications when used under conditions of proper protection and authorization and in accordance with all applicable legal and regulatory requirements. It refers to the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses.*
- ***"American leadership"*** *among nations worldwide will only be achieved by mobilizing the capabilities of all sectors of our greater society, including government at all levels, industry, foundations, academia, education, and individuals in using, supporting, and evolving the digital scientific data universe.*
- ***"Global information society"*** *recognizes that science and technology co-exist in a world where technology diminishes geographic, temporal, social, and national barriers to discovery, access, and use of data.*

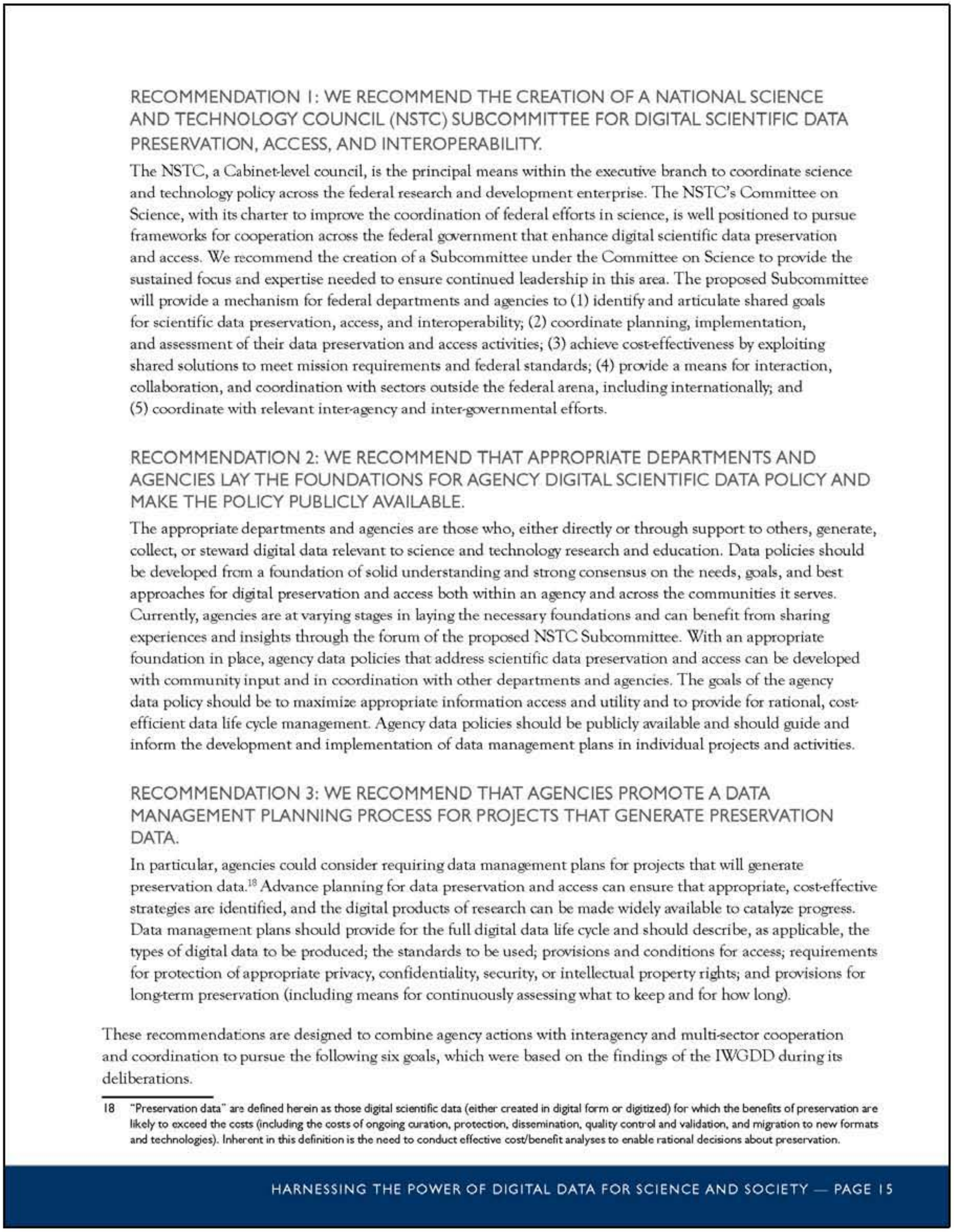
American leadership will only be achieved by mobilizing the capabilities of all sectors of our society.

This strategy is designed to unite the capabilities and leverage the resources of the federal agencies and organizations in their scientific data activities, thereby enabling the federal government to serve as both leader and partner to all sectors of our society in realizing the full potential of the digital dimension to enable discovery, innovation, and progress.

RECOMMENDATIONS: OUTLINE AND SUPPORTING GOALS

We make three recommendations pursuant to this strategy. These are presented at an outline level below, along with a discussion of the goals that support the recommendations. The final section of this report provides a more detailed discussion of the recommendations. The recommendations are intended to create synergy by combining action items for agencies to work with their communities of practice in pursuit of their respective missions. The recommendations also allow for a forum for cooperation and coordination across government, academic, commercial, and international sectors.

[17] Metadata are data about data. They include a formal description of the data, as well as information on how to acquire the data, and information for using the data, such as accuracy, security, and rights. Metadata provide the scientific, technical, contextual, representational, provenance, and other information necessary to enable creative re-use and re-purposing.



RECOMMENDATION 1: WE RECOMMEND THE CREATION OF A NATIONAL SCIENCE AND TECHNOLOGY COUNCIL (NSTC) SUBCOMMITTEE FOR DIGITAL SCIENTIFIC DATA PRESERVATION, ACCESS, AND INTEROPERABILITY.

The NSTC, a Cabinet-level council, is the principal means within the executive branch to coordinate science and technology policy across the federal research and development enterprise. The NSTC's Committee on Science, with its charter to improve the coordination of federal efforts in science, is well positioned to pursue frameworks for cooperation across the federal government that enhance digital scientific data preservation and access. We recommend the creation of a Subcommittee under the Committee on Science to provide the sustained focus and expertise needed to ensure continued leadership in this area. The proposed Subcommittee will provide a mechanism for federal departments and agencies to (1) identify and articulate shared goals for scientific data preservation, access, and interoperability; (2) coordinate planning, implementation, and assessment of their data preservation and access activities; (3) achieve cost-effectiveness by exploiting shared solutions to meet mission requirements and federal standards; (4) provide a means for interaction, collaboration, and coordination with sectors outside the federal arena, including internationally; and (5) coordinate with relevant inter-agency and inter-governmental efforts.

RECOMMENDATION 2: WE RECOMMEND THAT APPROPRIATE DEPARTMENTS AND AGENCIES LAY THE FOUNDATIONS FOR AGENCY DIGITAL SCIENTIFIC DATA POLICY AND MAKE THE POLICY PUBLICLY AVAILABLE.

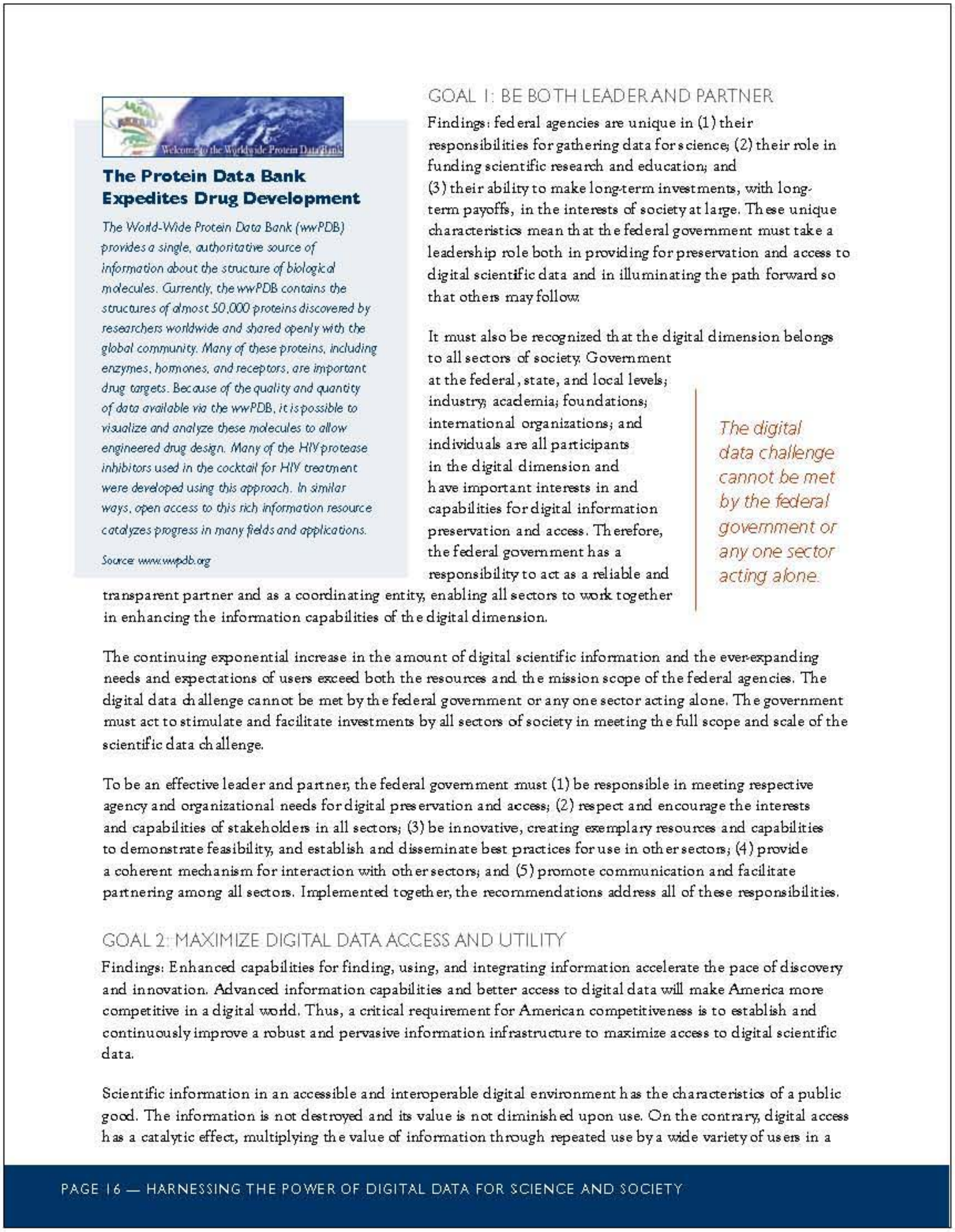
The appropriate departments and agencies are those who, either directly or through support to others, generate, collect, or steward digital data relevant to science and technology research and education. Data policies should be developed from a foundation of solid understanding and strong consensus on the needs, goals, and best approaches for digital preservation and access both within an agency and across the communities it serves. Currently, agencies are at varying stages in laying the necessary foundations and can benefit from sharing experiences and insights through the forum of the proposed NSTC Subcommittee. With an appropriate foundation in place, agency data policies that address scientific data preservation and access can be developed with community input and in coordination with other departments and agencies. The goals of the agency data policy should be to maximize appropriate information access and utility and to provide for rational, cost-efficient data life cycle management. Agency data policies should be publicly available and should guide and inform the development and implementation of data management plans in individual projects and activities.

RECOMMENDATION 3: WE RECOMMEND THAT AGENCIES PROMOTE A DATA MANAGEMENT PLANNING PROCESS FOR PROJECTS THAT GENERATE PRESERVATION DATA.

In particular, agencies could consider requiring data management plans for projects that will generate preservation data.[18] Advance planning for data preservation and access can ensure that appropriate, cost-effective strategies are identified, and the digital products of research can be made widely available to catalyze progress. Data management plans should provide for the full digital data life cycle and should describe, as applicable, the types of digital data to be produced; the standards to be used; provisions and conditions for access; requirements for protection of appropriate privacy, confidentiality, security, or intellectual property rights; and provisions for long-term preservation (including means for continuously assessing what to keep and for how long).

These recommendations are designed to combine agency actions with interagency and multi-sector cooperation and coordination to pursue the following six goals, which were based on the findings of the IWGDD during its deliberations.

[18] "Preservation data" are defined herein as those digital scientific data (either created in digital form or digitized) for which the benefits of preservation are likely to exceed the costs (including the costs of ongoing curation, protection, dissemination, quality control and validation, and migration to new formats and technologies). Inherent in this definition is the need to conduct effective cost/benefit analyses to enable rational decisions about preservation.



The Protein Data Bank Expedites Drug Development

The World-Wide Protein Data Bank (wwPDB) provides a single, authoritative source of information about the structure of biological molecules. Currently, the wwPDB contains the structures of almost 50,000 proteins discovered by researchers worldwide and shared openly with the global community. Many of these proteins, including enzymes, hormones, and receptors, are important drug targets. Because of the quality and quantity of data available via the wwPDB, it is possible to visualize and analyze these molecules to allow engineered drug design. Many of the HIV protease inhibitors used in the cocktail for HIV treatment were developed using this approach. In similar ways, open access to this rich information resource catalyzes progress in many fields and applications.

Source: www.wwpdb.org

GOAL 1: BE BOTH LEADER AND PARTNER

Findings: federal agencies are unique in (1) their responsibilities for gathering data for science; (2) their role in funding scientific research and education; and (3) their ability to make long-term investments, with long-term payoffs, in the interests of society at large. These unique characteristics mean that the federal government must take a leadership role both in providing for preservation and access to digital scientific data and in illuminating the path forward so that others may follow.

It must also be recognized that the digital dimension belongs to all sectors of society. Government at the federal, state, and local levels; industry; academia; foundations; international organizations; and individuals are all participants in the digital dimension and have important interests in and capabilities for digital information preservation and access. Therefore, the federal government has a responsibility to act as a reliable and transparent partner and as a coordinating entity, enabling all sectors to work together in enhancing the information capabilities of the digital dimension.

The digital data challenge cannot be met by the federal government or any one sector acting alone.

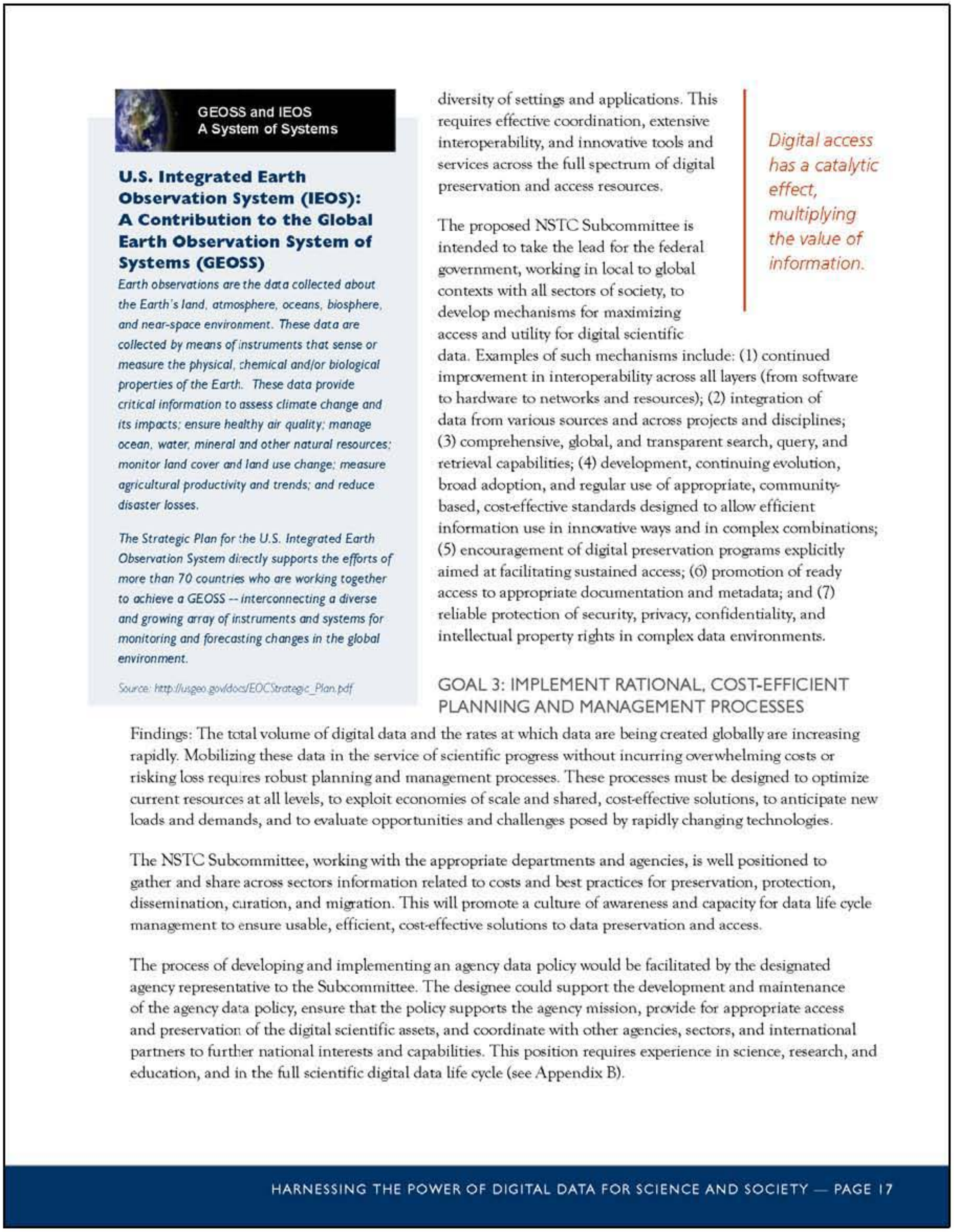
The continuing exponential increase in the amount of digital scientific information and the ever-expanding needs and expectations of users exceed both the resources and the mission scope of the federal agencies. The digital data challenge cannot be met by the federal government or any one sector acting alone. The government must act to stimulate and facilitate investments by all sectors of society in meeting the full scope and scale of the scientific data challenge.

To be an effective leader and partner, the federal government must (1) be responsible in meeting respective agency and organizational needs for digital preservation and access; (2) respect and encourage the interests and capabilities of stakeholders in all sectors; (3) be innovative, creating exemplary resources and capabilities to demonstrate feasibility, and establish and disseminate best practices for use in other sectors; (4) provide a coherent mechanism for interaction with other sectors; and (5) promote communication and facilitate partnering among all sectors. Implemented together, the recommendations address all of these responsibilities.

GOAL 2: MAXIMIZE DIGITAL DATA ACCESS AND UTILITY

Findings: Enhanced capabilities for finding, using, and integrating information accelerate the pace of discovery and innovation. Advanced information capabilities and better access to digital data will make America more competitive in a digital world. Thus, a critical requirement for American competitiveness is to establish and continuously improve a robust and pervasive information infrastructure to maximize access to digital scientific data.

Scientific information in an accessible and interoperable digital environment has the characteristics of a public good. The information is not destroyed and its value is not diminished upon use. On the contrary, digital access has a catalytic effect, multiplying the value of information through repeated use by a wide variety of users in a



GEOSS and IEOS
A System of Systems

U.S. Integrated Earth Observation System (IEOS): A Contribution to the Global Earth Observation System of Systems (GEOSS)

Earth observations are the data collected about the Earth's land, atmosphere, oceans, biosphere, and near-space environment. These data are collected by means of instruments that sense or measure the physical, chemical and/or biological properties of the Earth. These data provide critical information to assess climate change and its impacts; ensure healthy air quality; manage ocean, water, mineral and other natural resources; monitor land cover and land use change; measure agricultural productivity and trends; and reduce disaster losses.

The Strategic Plan for the U.S. Integrated Earth Observation System directly supports the efforts of more than 70 countries who are working together to achieve a GEOSS -- interconnecting a diverse and growing array of instruments and systems for monitoring and forecasting changes in the global environment.

Source: http://usgeo.gov/docs/EOCStrategic_Plan.pdf

diversity of settings and applications. This requires effective coordination, extensive interoperability, and innovative tools and services across the full spectrum of digital preservation and access resources.

The proposed NSTC Subcommittee is intended to take the lead for the federal government, working in local to global contexts with all sectors of society, to develop mechanisms for maximizing access and utility for digital scientific data. Examples of such mechanisms include: (1) continued improvement in interoperability across all layers (from software to hardware to networks and resources); (2) integration of data from various sources and across projects and disciplines; (3) comprehensive, global, and transparent search, query, and retrieval capabilities; (4) development, continuing evolution, broad adoption, and regular use of appropriate, community-based, cost-effective standards designed to allow efficient information use in innovative ways and in complex combinations; (5) encouragement of digital preservation programs explicitly aimed at facilitating sustained access; (6) promotion of ready access to appropriate documentation and metadata; and (7) reliable protection of security, privacy, confidentiality, and intellectual property rights in complex data environments.

Digital access has a catalytic effect, multiplying the value of information.

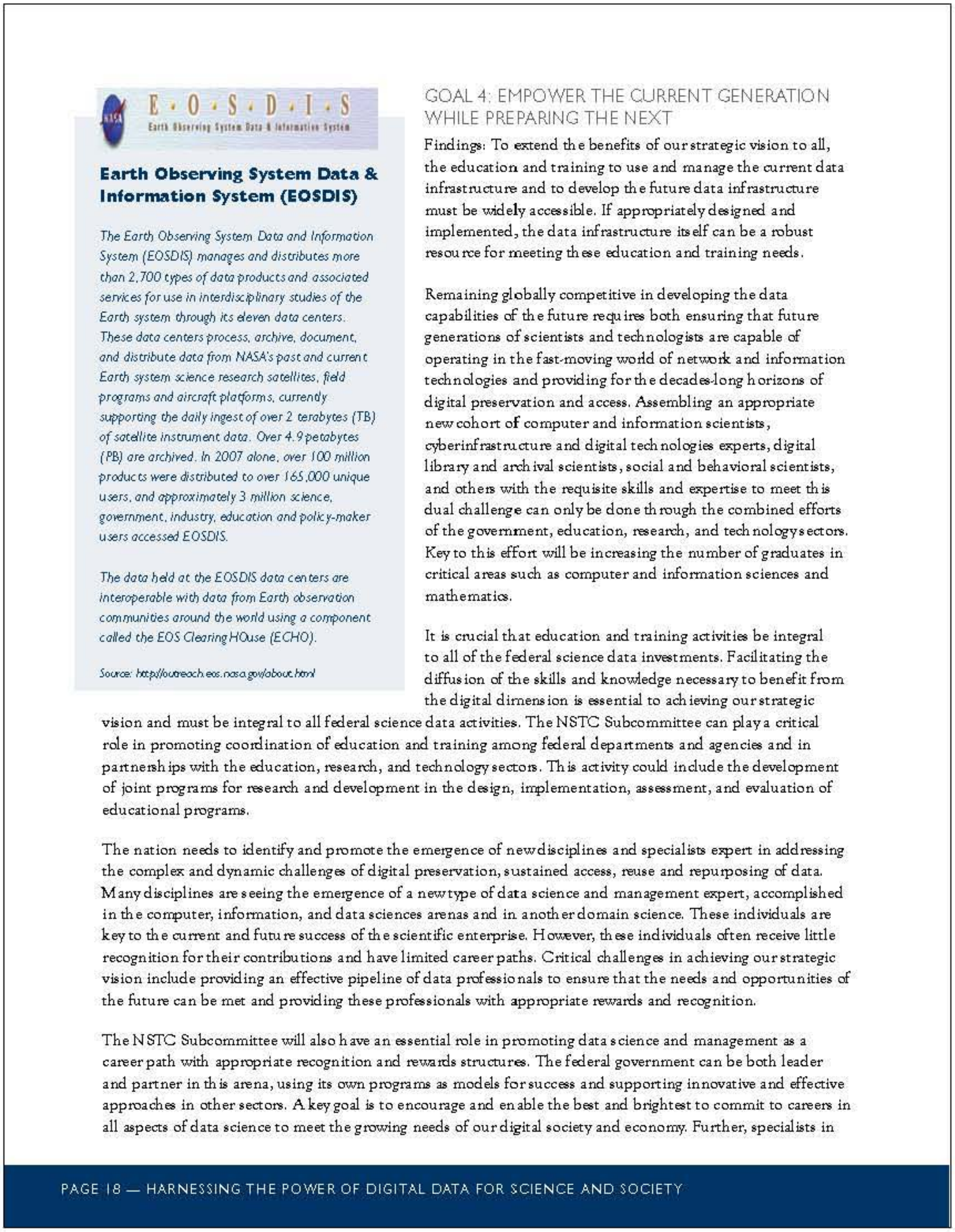
GOAL 3: IMPLEMENT RATIONAL, COST-EFFICIENT PLANNING AND MANAGEMENT PROCESSES

Findings: The total volume of digital data and the rates at which data are being created globally are increasing rapidly. Mobilizing these data in the service of scientific progress without incurring overwhelming costs or risking loss requires robust planning and management processes. These processes must be designed to optimize current resources at all levels, to exploit economies of scale and shared, cost-effective solutions, to anticipate new loads and demands, and to evaluate opportunities and challenges posed by rapidly changing technologies.

The NSTC Subcommittee, working with the appropriate departments and agencies, is well positioned to gather and share across sectors information related to costs and best practices for preservation, protection, dissemination, curation, and migration. This will promote a culture of awareness and capacity for data life cycle management to ensure usable, efficient, cost-effective solutions to data preservation and access.

The process of developing and implementing an agency data policy would be facilitated by the designated agency representative to the Subcommittee. The designee could support the development and maintenance of the agency data policy, ensure that the policy supports the agency mission, provide for appropriate access and preservation of the digital scientific assets, and coordinate with other agencies, sectors, and international partners to further national interests and capabilities. This position requires experience in science, research, and education, and in the full scientific digital data life cycle (see Appendix B).

HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY — PAGE 17



E · O · S · D · I · S

Earth Observing System Data & Information System

Earth Observing System Data & Information System (EOSDIS)

The Earth Observing System Data and Information System (EOSDIS) manages and distributes more than 2,700 types of data products and associated services for use in interdisciplinary studies of the Earth system through its eleven data centers. These data centers process, archive, document, and distribute data from NASA's past and current Earth system science research satellites, field programs and aircraft platforms, currently supporting the daily ingest of over 2 terabytes (TB) of satellite instrument data. Over 4.9 petabytes (PB) are archived. In 2007 alone, over 100 million products were distributed to over 165,000 unique users, and approximately 3 million science, government, industry, education and policy-maker users accessed EOSDIS.

The data held at the EOSDIS data centers are interoperable with data from Earth observation communities around the world using a component called the EOS Clearing HOuse (ECHO).

Source: http//outreach.eos.nasa.gov/about.html

GOAL 4: EMPOWER THE CURRENT GENERATION WHILE PREPARING THE NEXT

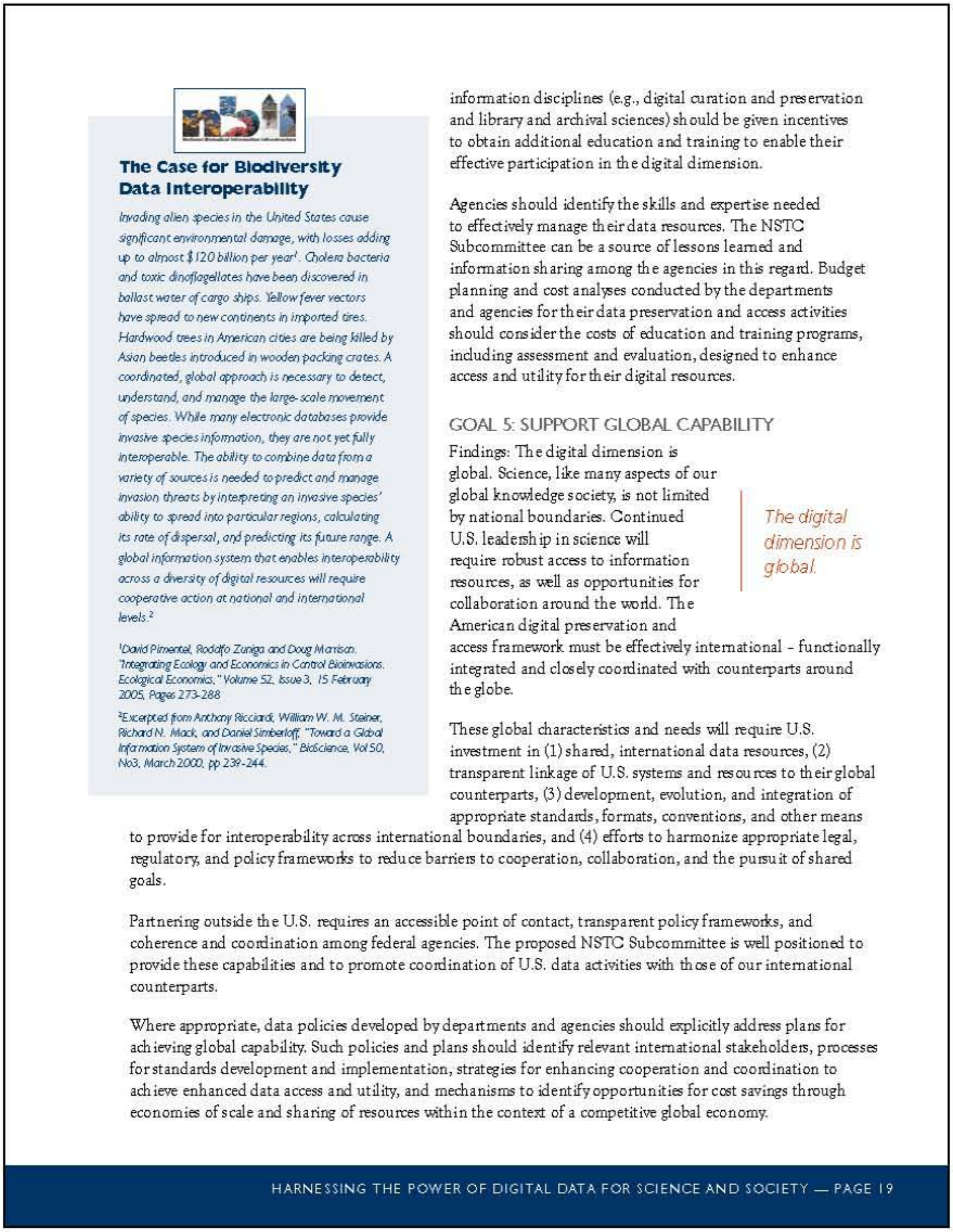
Findings: To extend the benefits of our strategic vision to all, the education and training to use and manage the current data infrastructure and to develop the future data infrastructure must be widely accessible. If appropriately designed and implemented, the data infrastructure itself can be a robust resource for meeting these education and training needs.

Remaining globally competitive in developing the data capabilities of the future requires both ensuring that future generations of scientists and technologists are capable of operating in the fast-moving world of network and information technologies and providing for the decades-long horizons of digital preservation and access. Assembling an appropriate new cohort of computer and information scientists, cyberinfrastructure and digital technologies experts, digital library and archival scientists, social and behavioral scientists, and others with the requisite skills and expertise to meet this dual challenge can only be done through the combined efforts of the government, education, research, and technology sectors. Key to this effort will be increasing the number of graduates in critical areas such as computer and information sciences and mathematics.

It is crucial that education and training activities be integral to all of the federal science data investments. Facilitating the diffusion of the skills and knowledge necessary to benefit from the digital dimension is essential to achieving our strategic vision and must be integral to all federal science data activities. The NSTC Subcommittee can play a critical role in promoting coordination of education and training among federal departments and agencies and in partnerships with the education, research, and technology sectors. This activity could include the development of joint programs for research and development in the design, implementation, assessment, and evaluation of educational programs.

The nation needs to identify and promote the emergence of new disciplines and specialists expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data. Many disciplines are seeing the emergence of a new type of data science and management expert, accomplished in the computer, information, and data sciences arenas and in another domain science. These individuals are key to the current and future success of the scientific enterprise. However, these individuals often receive little recognition for their contributions and have limited career paths. Critical challenges in achieving our strategic vision include providing an effective pipeline of data professionals to ensure that the needs and opportunities of the future can be met and providing these professionals with appropriate rewards and recognition.

The NSTC Subcommittee will also have an essential role in promoting data science and management as a career path with appropriate recognition and rewards structures. The federal government can be both leader and partner in this arena, using its own programs as models for success and supporting innovative and effective approaches in other sectors. A key goal is to encourage and enable the best and brightest to commit to careers in all aspects of data science to meet the growing needs of our digital society and economy. Further, specialists in



information disciplines (e.g., digital curation and preservation and library and archival sciences) should be given incentives to obtain additional education and training to enable their effective participation in the digital dimension.

Agencies should identify the skills and expertise needed to effectively manage their data resources. The NSTC Subcommittee can be a source of lessons learned and information sharing among the agencies in this regard. Budget planning and cost analyses conducted by the departments and agencies for their data preservation and access activities should consider the costs of education and training programs, including assessment and evaluation, designed to enhance access and utility for their digital resources.

GOAL 5: SUPPORT GLOBAL CAPABILITY

Findings: The digital dimension is global. Science, like many aspects of our global knowledge society, is not limited by national boundaries. Continued U.S. leadership in science will require robust access to information resources, as well as opportunities for collaboration around the world. The American digital preservation and access framework must be effectively international – functionally integrated and closely coordinated with counterparts around the globe.

The digital dimension is global.

These global characteristics and needs will require U.S. investment in (1) shared, international data resources, (2) transparent linkage of U.S. systems and resources to their global counterparts, (3) development, evolution, and integration of appropriate standards, formats, conventions, and other means to provide for interoperability across international boundaries, and (4) efforts to harmonize appropriate legal, regulatory, and policy frameworks to reduce barriers to cooperation, collaboration, and the pursuit of shared goals.

Partnering outside the U.S. requires an accessible point of contact, transparent policy frameworks, and coherence and coordination among federal agencies. The proposed NSTC Subcommittee is well positioned to provide these capabilities and to promote coordination of U.S. data activities with those of our international counterparts.

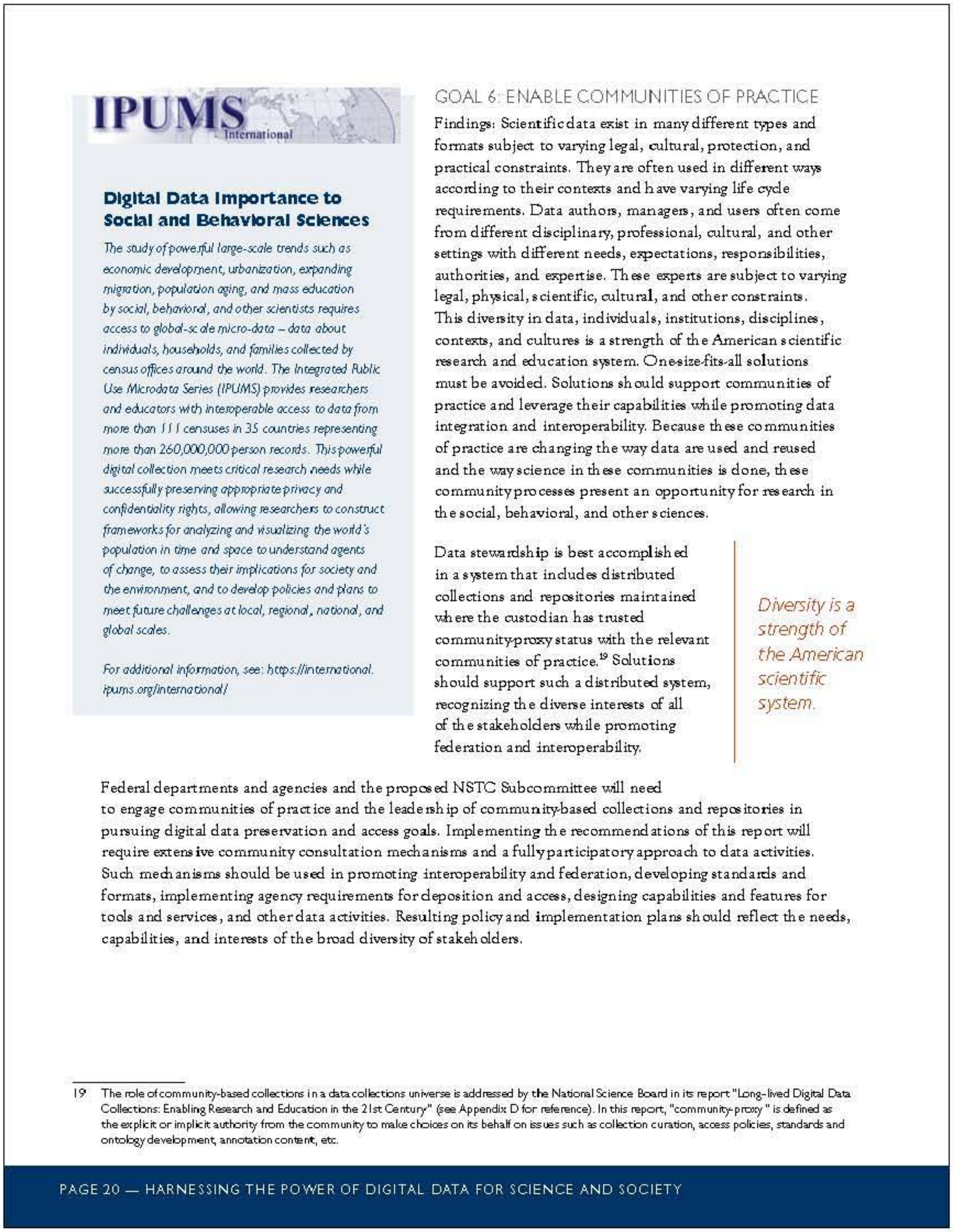
Where appropriate, data policies developed by departments and agencies should explicitly address plans for achieving global capability. Such policies and plans should identify relevant international stakeholders, processes for standards development and implementation, strategies for enhancing cooperation and coordination to achieve enhanced data access and utility, and mechanisms to identify opportunities for cost savings through economies of scale and sharing of resources within the context of a competitive global economy.

The Case for Biodiversity Data Interoperability

Invading alien species in the United States cause significant environmental damage, with losses adding up to almost $120 billion per year[1]. Cholera bacteria and toxic dinoflagellates have been discovered in ballast water of cargo ships. Yellow fever vectors have spread to new continents in imported tires. Hardwood trees in American cities are being killed by Asian beetles introduced in wooden packing crates. A coordinated, global approach is necessary to detect, understand, and manage the large-scale movement of species. While many electronic databases provide invasive species information, they are not yet fully interoperable. The ability to combine data from a variety of sources is needed to predict and manage invasion threats by interpreting an invasive species' ability to spread into particular regions, calculating its rate of dispersal, and predicting its future range. A global information system that enables interoperability across a diversity of digital resources will require cooperative action at national and international levels.[2]

[1] David Pimentel, Rodolfo Zuniga and Doug Morrison. "Integrating Ecology and Economics in Control Bioinvasions. Ecological Economics," Volume 52, Issue 3, 15 February 2005, Pages 273-288

[2] Excerpted from Anthony Ricciardi, William W. M. Steiner, Richard N. Mack, and Daniel Simberloff, "Toward a Global Information System of Invasive Species," BioScience, Vol 50, No3, March 2000, pp 239-244.



IPUMS International

Digital Data Importance to Social and Behavioral Sciences

The study of powerful large-scale trends such as economic development, urbanization, expanding migration, population aging, and mass education by social, behavioral, and other scientists requires access to global-scale micro-data – data about individuals, households, and families collected by census offices around the world. The Integrated Public Use Microdata Series (IPUMS) provides researchers and educators with interoperable access to data from more than 111 censuses in 35 countries representing more than 260,000,000 person records. This powerful digital collection meets critical research needs while successfully preserving appropriate privacy and confidentiality rights, allowing researchers to construct frameworks for analyzing and visualizing the world's population in time and space to understand agents of change, to assess their implications for society and the environment, and to develop policies and plans to meet future challenges at local, regional, national, and global scales.

For additional information, see: https://international.ipums.org/international/

GOAL 6: ENABLE COMMUNITIES OF PRACTICE

Findings: Scientific data exist in many different types and formats subject to varying legal, cultural, protection, and practical constraints. They are often used in different ways according to their contexts and have varying life cycle requirements. Data authors, managers, and users often come from different disciplinary, professional, cultural, and other settings with different needs, expectations, responsibilities, authorities, and expertise. These experts are subject to varying legal, physical, scientific, cultural, and other constraints. This diversity in data, individuals, institutions, disciplines, contexts, and cultures is a strength of the American scientific research and education system. One-size-fits-all solutions must be avoided. Solutions should support communities of practice and leverage their capabilities while promoting data integration and interoperability. Because these communities of practice are changing the way data are used and reused and the way science in these communities is done, these community processes present an opportunity for research in the social, behavioral, and other sciences.

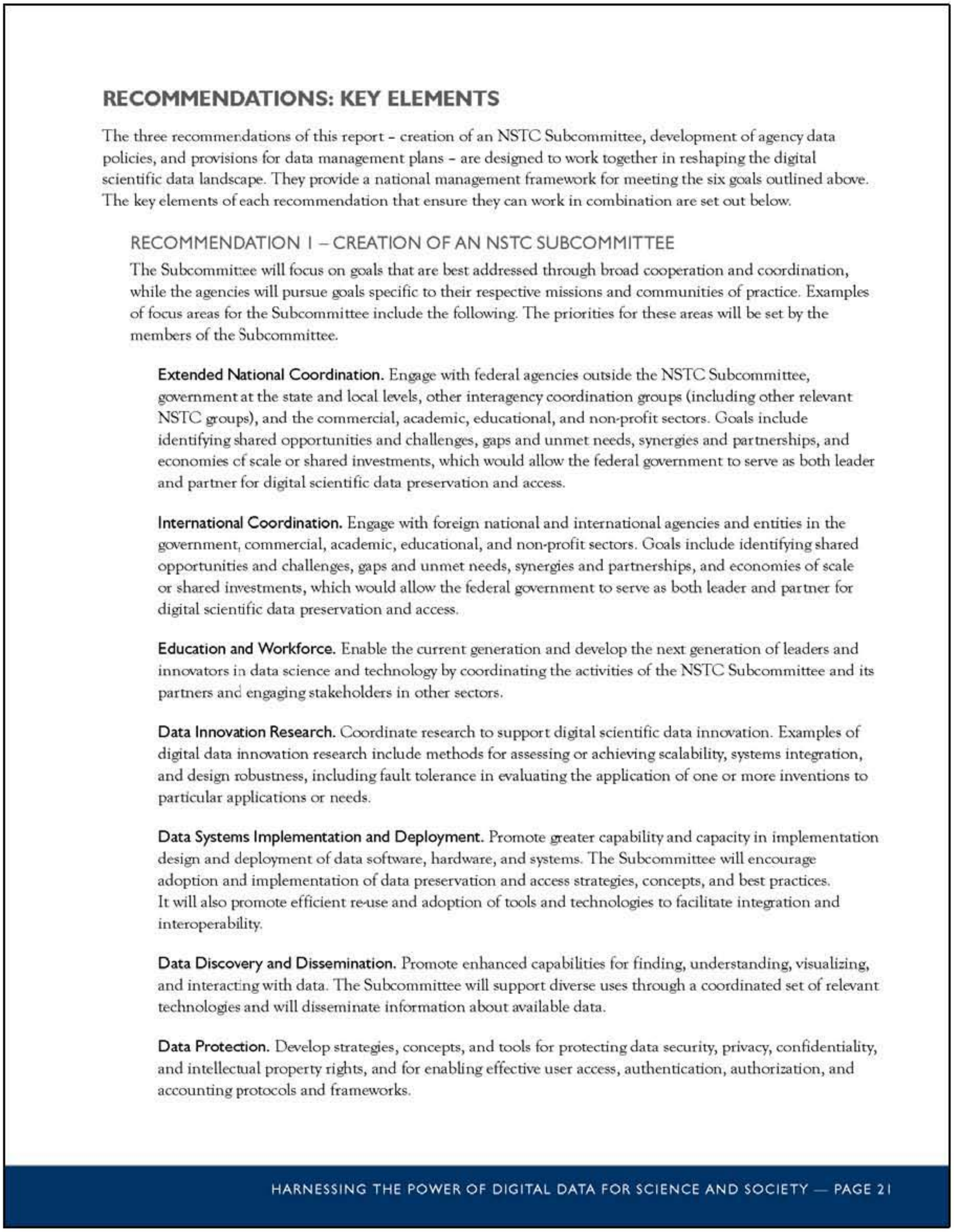
Data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice.[19] Solutions should support such a distributed system, recognizing the diverse interests of all of the stakeholders while promoting federation and interoperability.

Diversity is a strength of the American scientific system.

Federal departments and agencies and the proposed NSTC Subcommittee will need to engage communities of practice and the leadership of community-based collections and repositories in pursuing digital data preservation and access goals. Implementing the recommendations of this report will require extensive community consultation mechanisms and a fully participatory approach to data activities. Such mechanisms should be used in promoting interoperability and federation, developing standards and formats, implementing agency requirements for deposition and access, designing capabilities and features for tools and services, and other data activities. Resulting policy and implementation plans should reflect the needs, capabilities, and interests of the broad diversity of stakeholders.

[19] The role of community-based collections in a data collections universe is addressed by the National Science Board in its report "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" (see Appendix D for reference). In this report, "community-proxy " is defined as the explicit or implicit authority from the community to make choices on its behalf on issues such as collection curation, access policies, standards and ontology development, annotation content, etc.

PAGE 20 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



RECOMMENDATIONS: KEY ELEMENTS

The three recommendations of this report – creation of an NSTC Subcommittee, development of agency data policies, and provisions for data management plans – are designed to work together in reshaping the digital scientific data landscape. They provide a national management framework for meeting the six goals outlined above. The key elements of each recommendation that ensure they can work in combination are set out below.

RECOMMENDATION 1 – CREATION OF AN NSTC SUBCOMMITTEE

The Subcommittee will focus on goals that are best addressed through broad cooperation and coordination, while the agencies will pursue goals specific to their respective missions and communities of practice. Examples of focus areas for the Subcommittee include the following. The priorities for these areas will be set by the members of the Subcommittee.

Extended National Coordination. Engage with federal agencies outside the NSTC Subcommittee, government at the state and local levels, other interagency coordination groups (including other relevant NSTC groups), and the commercial, academic, educational, and non-profit sectors. Goals include identifying shared opportunities and challenges, gaps and unmet needs, synergies and partnerships, and economies cf scale or shared investments, which would allow the federal government to serve as both leader and partner for digital scientific data preservation and access.

International Coordination. Engage with foreign national and international agencies and entities in the government, commercial, academic, educational, and non-profit sectors. Goals include identifying shared opportunities and challenges, gaps and unmet needs, synergies and partnerships, and economies of scale or shared investments, which would allow the federal government to serve as both leader and partner for digital scientific data preservation and access.

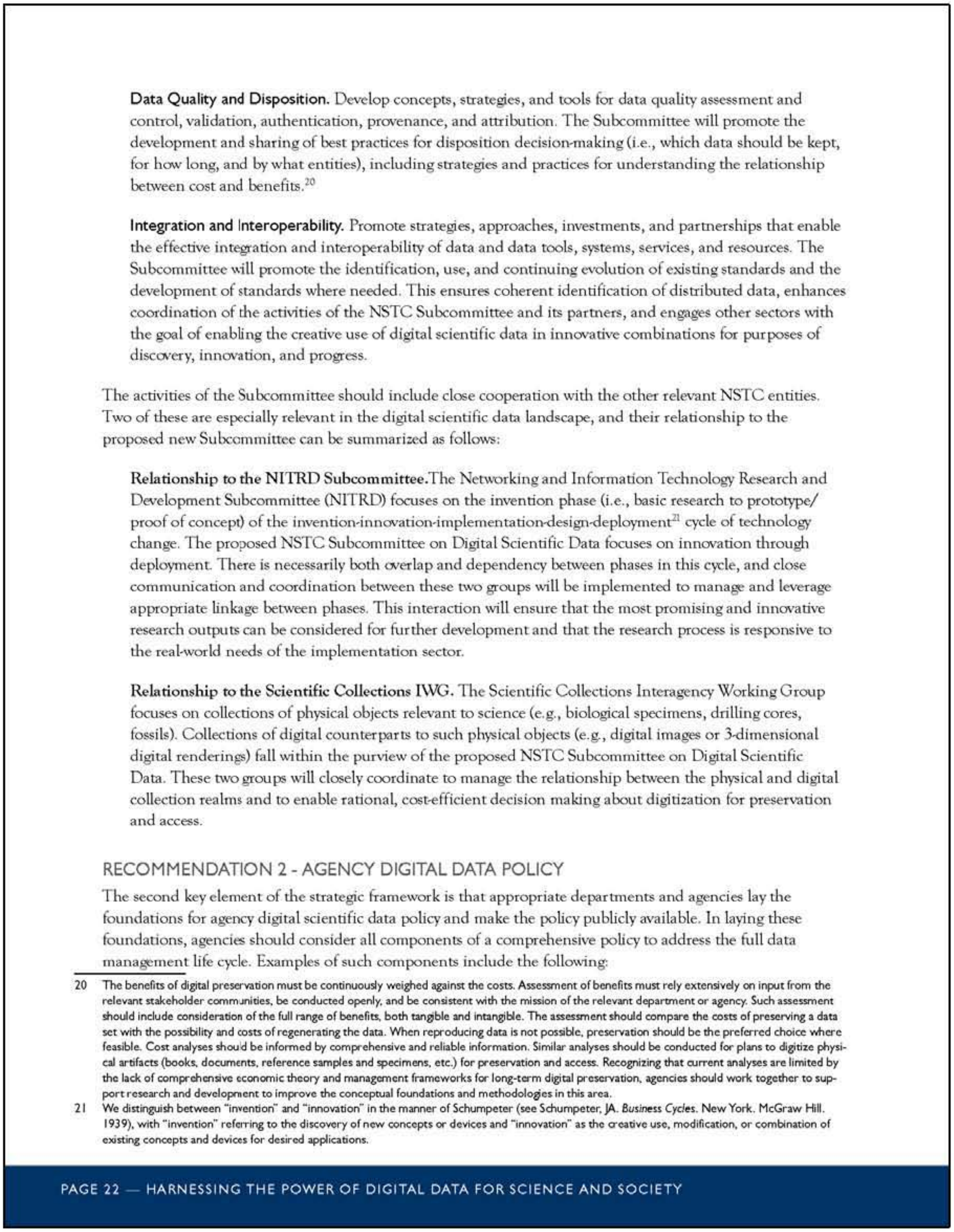
Education and Workforce. Enable the current generation and develop the next generation of leaders and innovators in data science and technology by coordinating the activities of the NSTC Subcommittee and its partners and engaging stakeholders in other sectors.

Data Innovation Research. Coordinate research to support digital scientific data innovation. Examples of digital data innovation research include methods for assessing or achieving scalability, systems integration, and design robustness, including fault tolerance in evaluating the application of one or more inventions to particular applications or needs.

Data Systems Implementation and Deployment. Promote greater capability and capacity in implementation design and deployment of data software, hardware, and systems. The Subcommittee will encourage adoption and implementation of data preservation and access strategies, concepts, and best practices. It will also promote efficient re-use and adoption of tools and technologies to facilitate integration and interoperability.

Data Discovery and Dissemination. Promote enhanced capabilities for finding, understanding, visualizing, and interacting with data. The Subcommittee will support diverse uses through a coordinated set of relevant technologies and will disseminate information about available data.

Data Protection. Develop strategies, concepts, and tools for protecting data security, privacy, confidentiality, and intellectual property rights, and for enabling effective user access, authentication, authorization, and accounting protocols and frameworks.



Data Quality and Disposition. Develop concepts, strategies, and tools for data quality assessment and control, validation, authentication, provenance, and attribution. The Subcommittee will promote the development and sharing of best practices for disposition decision-making (i.e., which data should be kept, for how long, and by what entities), including strategies and practices for understanding the relationship between cost and benefits.[20]

Integration and Interoperability. Promote strategies, approaches, investments, and partnerships that enable the effective integration and interoperability of data and data tools, systems, services, and resources. The Subcommittee will promote the identification, use, and continuing evolution of existing standards and the development of standards where needed. This ensures coherent identification of distributed data, enhances coordination of the activities of the NSTC Subcommittee and its partners, and engages other sectors with the goal of enabling the creative use of digital scientific data in innovative combinations for purposes of discovery, innovation, and progress.

The activities of the Subcommittee should include close cooperation with the other relevant NSTC entities. Two of these are especially relevant in the digital scientific data landscape, and their relationship to the proposed new Subcommittee can be summarized as follows:

Relationship to the NITRD Subcommittee. The Networking and Information Technology Research and Development Subcommittee (NITRD) focuses on the invention phase (i.e., basic research to prototype/ proof of concept) of the invention-innovation-implementation-design-deployment[21] cycle of technology change. The proposed NSTC Subcommittee on Digital Scientific Data focuses on innovation through deployment. There is necessarily both overlap and dependency between phases in this cycle, and close communication and coordination between these two groups will be implemented to manage and leverage appropriate linkage between phases. This interaction will ensure that the most promising and innovative research outputs can be considered for further development and that the research process is responsive to the real-world needs of the implementation sector.

Relationship to the Scientific Collections IWG. The Scientific Collections Interagency Working Group focuses on collections of physical objects relevant to science (e.g., biological specimens, drilling cores, fossils). Collections of digital counterparts to such physical objects (e.g., digital images or 3-dimensional digital renderings) fall within the purview of the proposed NSTC Subcommittee on Digital Scientific Data. These two groups will closely coordinate to manage the relationship between the physical and digital collection realms and to enable rational, cost-efficient decision making about digitization for preservation and access.

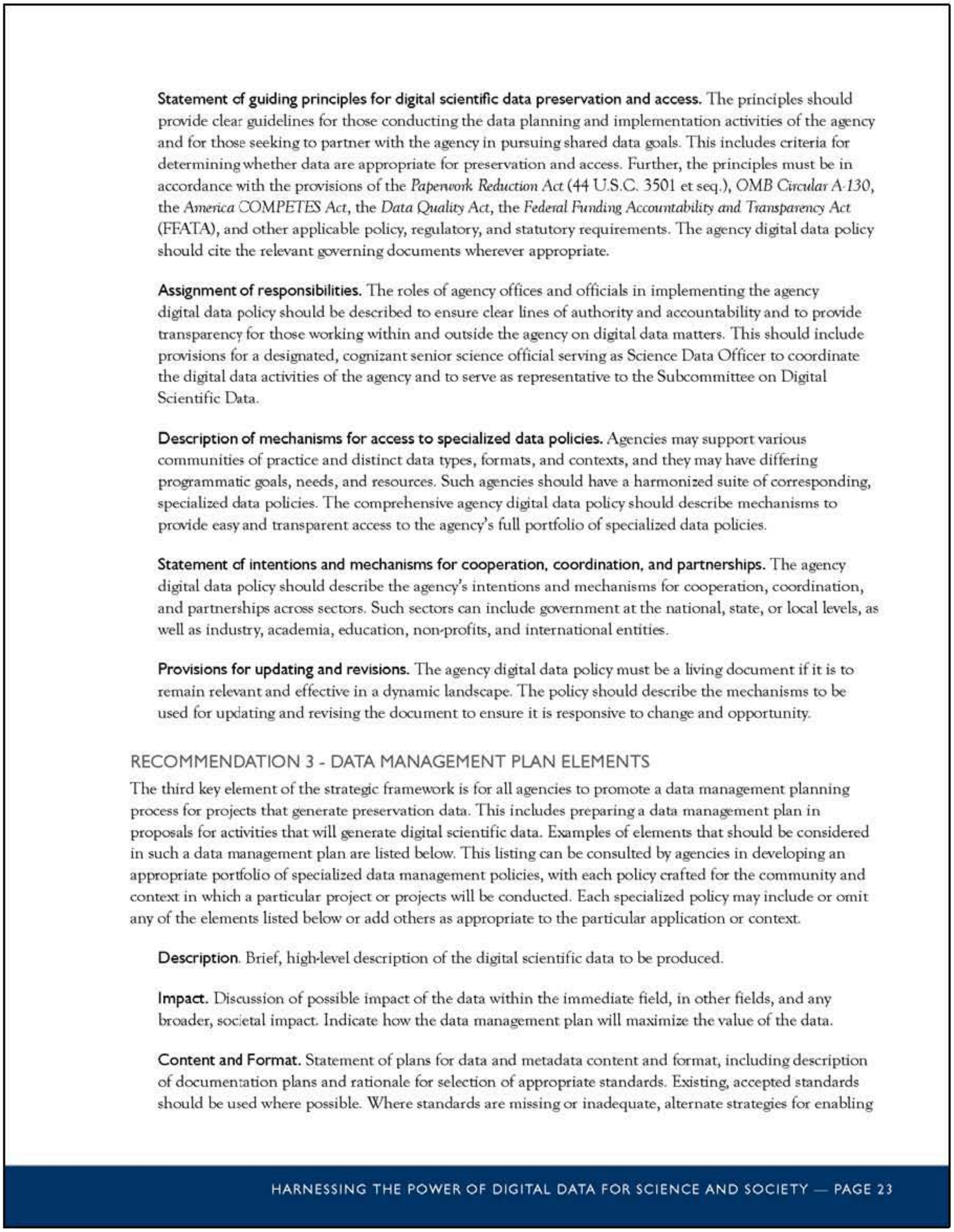
RECOMMENDATION 2 - AGENCY DIGITAL DATA POLICY

The second key element of the strategic framework is that appropriate departments and agencies lay the foundations for agency digital scientific data policy and make the policy publicly available. In laying these foundations, agencies should consider all components of a comprehensive policy to address the full data management life cycle. Examples of such components include the following:

[20] The benefits of digital preservation must be continuously weighed against the costs. Assessment of benefits must rely extensively on input from the relevant stakeholder communities, be conducted openly, and be consistent with the mission of the relevant department or agency. Such assessment should include consideration of the full range of benefits, both tangible and intangible. The assessment should compare the costs of preserving a data set with the possibility and costs of regenerating the data. When reproducing data is not possible, preservation should be the preferred choice where feasible. Cost analyses should be informed by comprehensive and reliable information. Similar analyses should be conducted for plans to digitize physical artifacts (books, documents, reference samples and specimens, etc.) for preservation and access. Recognizing that current analyses are limited by the lack of comprehensive economic theory and management frameworks for long-term digital preservation, agencies should work together to support research and development to improve the conceptual foundations and methodologies in this area.

[21] We distinguish between "invention" and "innovation" in the manner of Schumpeter (see Schumpeter, JA. *Business Cycles*. New York. McGraw Hill. 1939), with "invention" referring to the discovery of new concepts or devices and "innovation" as the creative use, modification, or combination of existing concepts and devices for desired applications.

PAGE 22 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



Statement of guiding principles for digital scientific data preservation and access. The principles should provide clear guidelines for those conducting the data planning and implementation activities of the agency and for those seeking to partner with the agency in pursuing shared data goals. This includes criteria for determining whether data are appropriate for preservation and access. Further, the principles must be in accordance with the provisions of the *Paperwork Reduction Act* (44 U.S.C. 3501 et seq.), *OMB Circular A-130*, the *America COMPETES Act*, the *Data Quality Act*, the *Federal Funding Accountability and Transparency Act* (FFATA), and other applicable policy, regulatory, and statutory requirements. The agency digital data policy should cite the relevant governing documents wherever appropriate.

Assignment of responsibilities. The roles of agency offices and officials in implementing the agency digital data policy should be described to ensure clear lines of authority and accountability and to provide transparency for those working within and outside the agency on digital data matters. This should include provisions for a designated, cognizant senior science official serving as Science Data Officer to coordinate the digital data activities of the agency and to serve as representative to the Subcommittee on Digital Scientific Data.

Description of mechanisms for access to specialized data policies. Agencies may support various communities of practice and distinct data types, formats, and contexts, and they may have differing programmatic goals, needs, and resources. Such agencies should have a harmonized suite of corresponding, specialized data policies. The comprehensive agency digital data policy should describe mechanisms to provide easy and transparent access to the agency's full portfolio of specialized data policies.

Statement of intentions and mechanisms for cooperation, coordination, and partnerships. The agency digital data policy should describe the agency's intentions and mechanisms for cooperation, coordination, and partnerships across sectors. Such sectors can include government at the national, state, or local levels, as well as industry, academia, education, non-profits, and international entities.

Provisions for updating and revisions. The agency digital data policy must be a living document if it is to remain relevant and effective in a dynamic landscape. The policy should describe the mechanisms to be used for updating and revising the document to ensure it is responsive to change and opportunity.

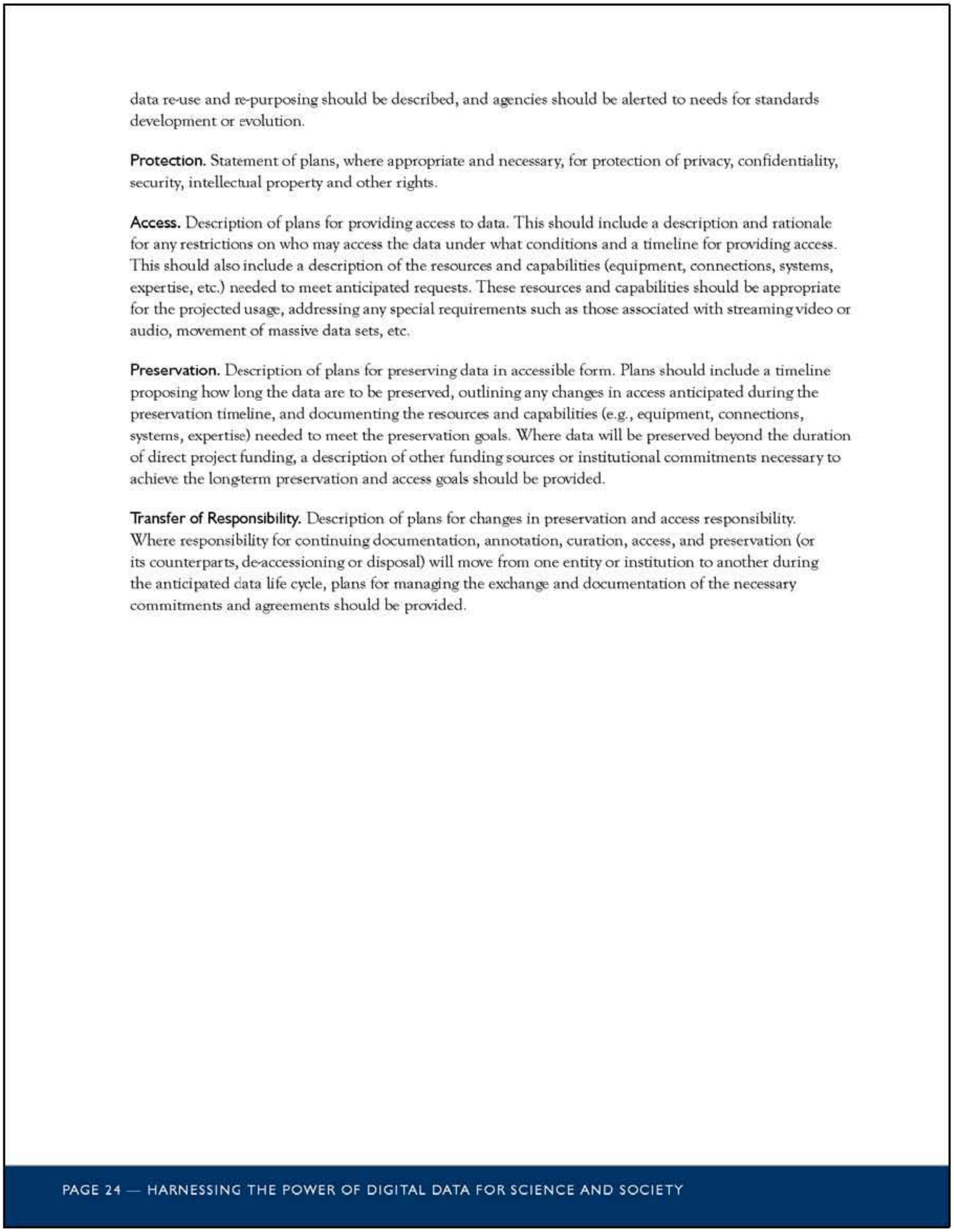
RECOMMENDATION 3 - DATA MANAGEMENT PLAN ELEMENTS

The third key element of the strategic framework is for all agencies to promote a data management planning process for projects that generate preservation data. This includes preparing a data management plan in proposals for activities that will generate digital scientific data. Examples of elements that should be considered in such a data management plan are listed below. This listing can be consulted by agencies in developing an appropriate portfolio of specialized data management policies, with each policy crafted for the community and context in which a particular project or projects will be conducted. Each specialized policy may include or omit any of the elements listed below or add others as appropriate to the particular application or context.

Description. Brief, high-level description of the digital scientific data to be produced.

Impact. Discussion of possible impact of the data within the immediate field, in other fields, and any broader, societal impact. Indicate how the data management plan will maximize the value of the data.

Content and Format. Statement of plans for data and metadata content and format, including description of documentation plans and rationale for selection of appropriate standards. Existing, accepted standards should be used where possible. Where standards are missing or inadequate, alternate strategies for enabling



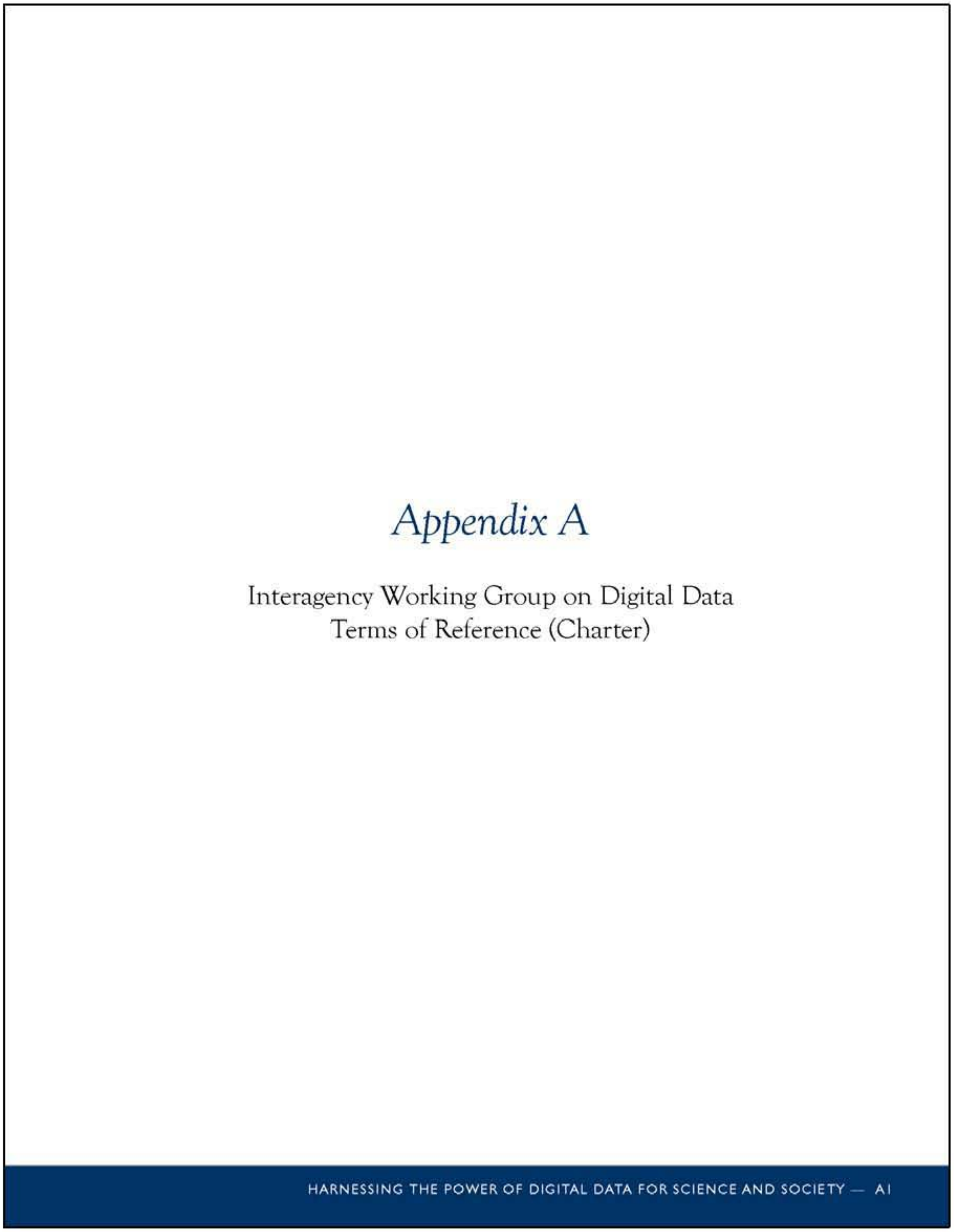
data re-use and re-purposing should be described, and agencies should be alerted to needs for standards development or evolution.

Protection. Statement of plans, where appropriate and necessary, for protection of privacy, confidentiality, security, intellectual property and other rights.

Access. Description of plans for providing access to data. This should include a description and rationale for any restrictions on who may access the data under what conditions and a timeline for providing access. This should also include a description of the resources and capabilities (equipment, connections, systems, expertise, etc.) needed to meet anticipated requests. These resources and capabilities should be appropriate for the projected usage, addressing any special requirements such as those associated with streaming video or audio, movement of massive data sets, etc.

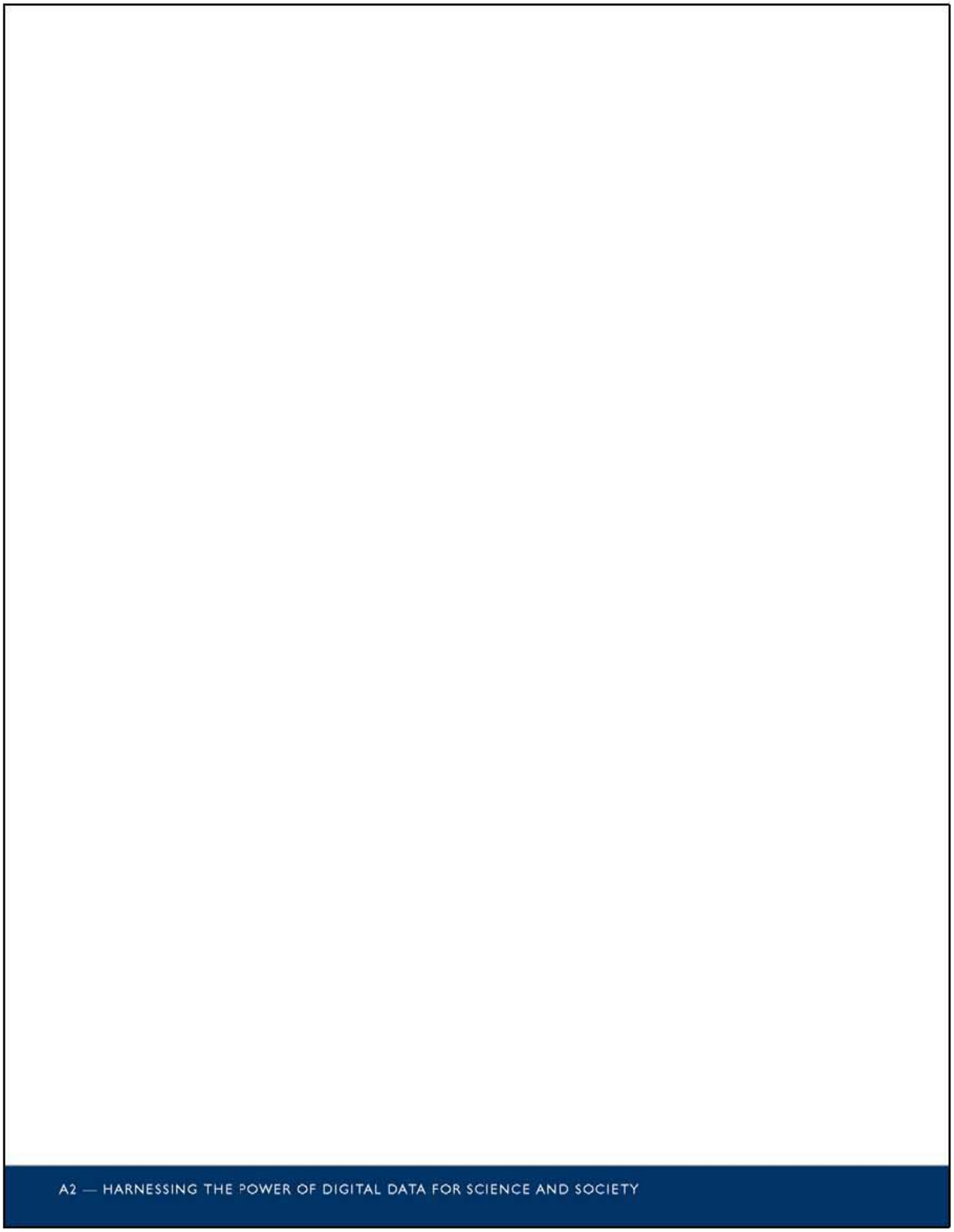
Preservation. Description of plans for preserving data in accessible form. Plans should include a timeline proposing how long the data are to be preserved, outlining any changes in access anticipated during the preservation timeline, and documenting the resources and capabilities (e.g., equipment, connections, systems, expertise) needed to meet the preservation goals. Where data will be preserved beyond the duration of direct project funding, a description of other funding sources or institutional commitments necessary to achieve the long-term preservation and access goals should be provided.

Transfer of Responsibility. Description of plans for changes in preservation and access responsibility. Where responsibility for continuing documentation, annotation, curation, access, and preservation (or its counterparts, de-accessioning or disposal) will move from one entity or institution to another during the anticipated data life cycle, plans for managing the exchange and documentation of the necessary commitments and agreements should be provided.

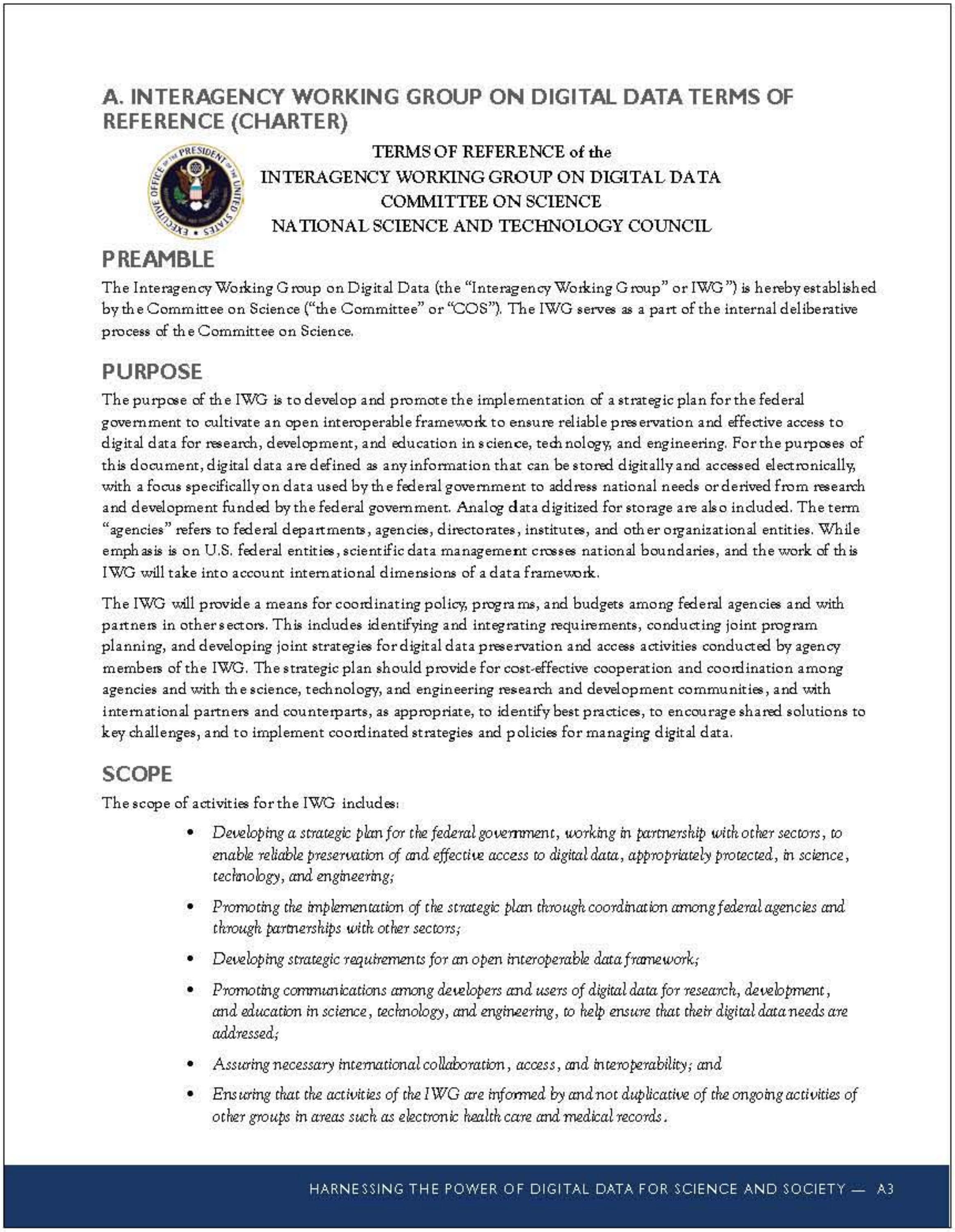


Appendix A

Interagency Working Group on Digital Data
Terms of Reference (Charter)



A2 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



A. INTERAGENCY WORKING GROUP ON DIGITAL DATA TERMS OF REFERENCE (CHARTER)

TERMS OF REFERENCE of the
INTERAGENCY WORKING GROUP ON DIGITAL DATA
COMMITTEE ON SCIENCE
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

PREAMBLE

The Interagency Working Group on Digital Data (the "Interagency Working Group" or IWG") is hereby established by the Committee on Science ("the Committee" or "COS"). The IWG serves as a part of the internal deliberative process of the Committee on Science.

PURPOSE

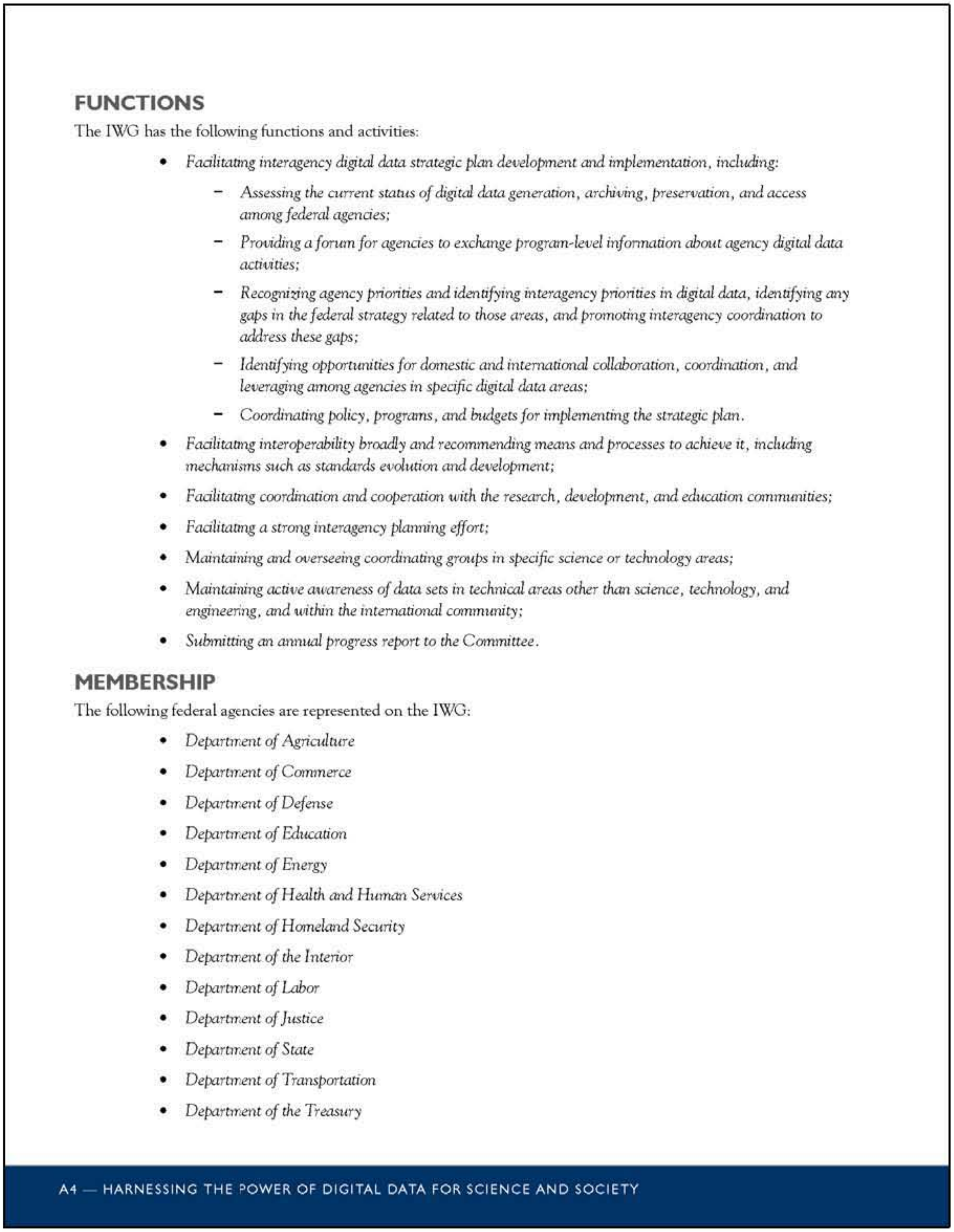
The purpose of the IWG is to develop and promote the implementation of a strategic plan for the federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering. For the purposes of this document, digital data are defined as any information that can be stored digitally and accessed electronically, with a focus specifically on data used by the federal government to address national needs or derived from research and development funded by the federal government. Analog data digitized for storage are also included. The term "agencies" refers to federal departments, agencies, directorates, institutes, and other organizational entities. While emphasis is on U.S. federal entities, scientific data management crosses national boundaries, and the work of this IWG will take into account international dimensions of a data framework.

The IWG will provide a means for coordinating policy, programs, and budgets among federal agencies and with partners in other sectors. This includes identifying and integrating requirements, conducting joint program planning, and developing joint strategies for digital data preservation and access activities conducted by agency members of the IWG. The strategic plan should provide for cost-effective cooperation and coordination among agencies and with the science, technology, and engineering research and development communities, and with international partners and counterparts, as appropriate, to identify best practices, to encourage shared solutions to key challenges, and to implement coordinated strategies and policies for managing digital data.

SCOPE

The scope of activities for the IWG includes:

- *Developing a strategic plan for the federal government, working in partnership with other sectors, to enable reliable preservation of and effective access to digital data, appropriately protected, in science, technology, and engineering;*
- *Promoting the implementation of the strategic plan through coordination among federal agencies and through partnerships with other sectors;*
- *Developing strategic requirements for an open interoperable data framework;*
- *Promoting communications among developers and users of digital data for research, development, and education in science, technology, and engineering, to help ensure that their digital data needs are addressed;*
- *Assuring necessary international collaboration, access, and interoperability; and*
- *Ensuring that the activities of the IWG are informed by and not duplicative of the ongoing activities of other groups in areas such as electronic health care and medical records.*



FUNCTIONS

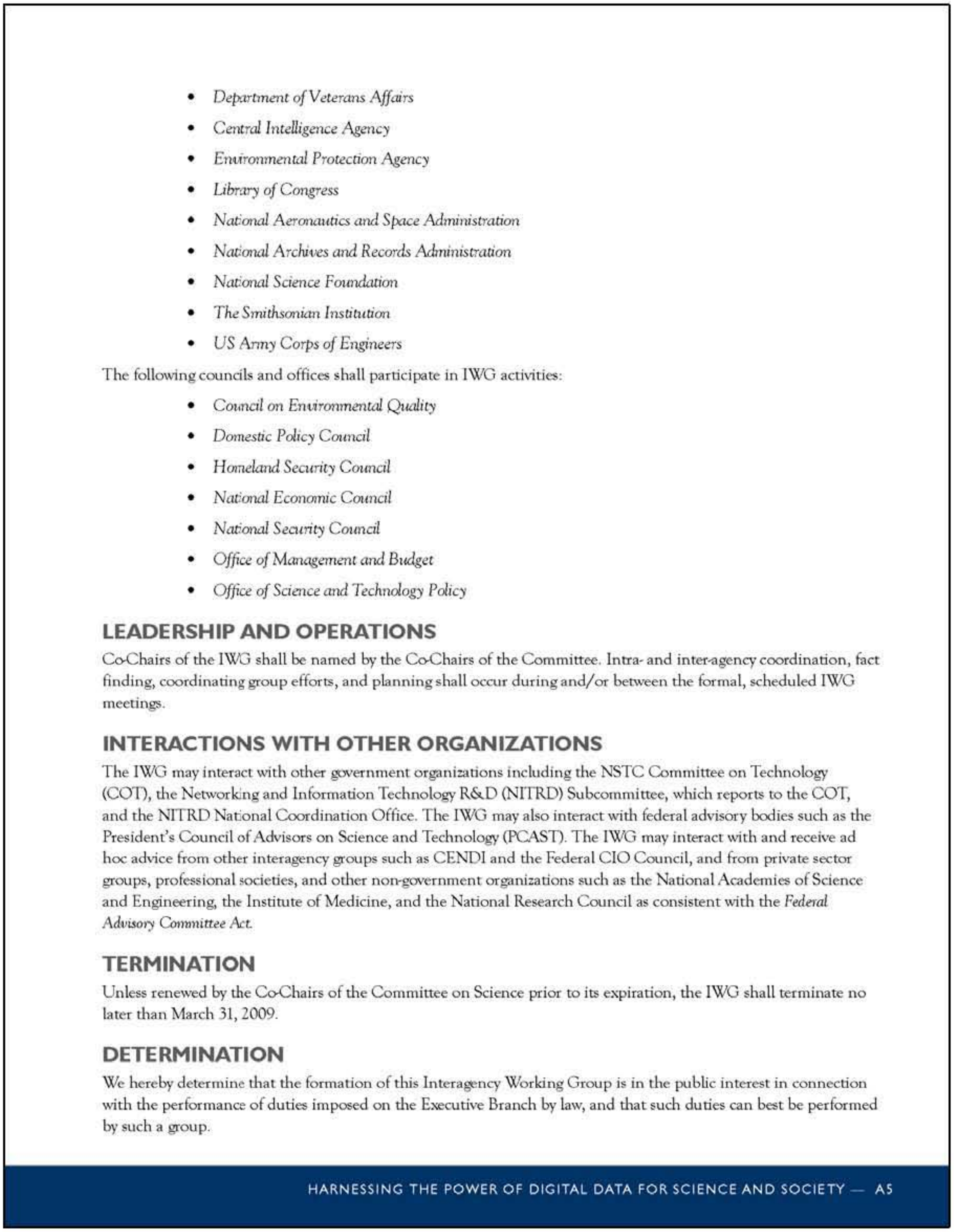
The IWG has the following functions and activities:

- *Facilitating interagency digital data strategic plan development and implementation, including:*
  - *Assessing the current status of digital data generation, archiving, preservation, and access among federal agencies;*
  - *Providing a forum for agencies to exchange program-level information about agency digital data activities;*
  - *Recognizing agency priorities and identifying interagency priorities in digital data, identifying any gaps in the federal strategy related to those areas, and promoting interagency coordination to address these gaps;*
  - *Identifying opportunities for domestic and international collaboration, coordination, and leveraging among agencies in specific digital data areas;*
  - *Coordinating policy, programs, and budgets for implementing the strategic plan.*
- *Facilitating interoperability broadly and recommending means and processes to achieve it, including mechanisms such as standards evolution and development;*
- *Facilitating coordination and cooperation with the research, development, and education communities;*
- *Facilitating a strong interagency planning effort;*
- *Maintaining and overseeing coordinating groups in specific science or technology areas;*
- *Maintaining active awareness of data sets in technical areas other than science, technology, and engineering, and within the international community;*
- *Submitting an annual progress report to the Committee.*

MEMBERSHIP

The following federal agencies are represented on the IWG:

- *Department of Agriculture*
- *Department of Commerce*
- *Department of Defense*
- *Department of Education*
- *Department of Energy*
- *Department of Health and Human Services*
- *Department of Homeland Security*
- *Department of the Interior*
- *Department of Labor*
- *Department of Justice*
- *Department of State*
- *Department of Transportation*
- *Department of the Treasury*



- *Department of Veterans Affairs*
- *Central Intelligence Agency*
- *Environmental Protection Agency*
- *Library of Congress*
- *National Aeronautics and Space Administration*
- *National Archives and Records Administration*
- *National Science Foundation*
- *The Smithsonian Institution*
- *US Army Corps of Engineers*

The following councils and offices shall participate in IWG activities:

- *Council on Environmental Quality*
- *Domestic Policy Council*
- *Homeland Security Council*
- *National Economic Council*
- *National Security Council*
- *Office of Management and Budget*
- *Office of Science and Technology Policy*

LEADERSHIP AND OPERATIONS

Co-Chairs of the IWG shall be named by the Co-Chairs of the Committee. Intra- and inter-agency coordination, fact finding, coordinating group efforts, and planning shall occur during and/or between the formal, scheduled IWG meetings.

INTERACTIONS WITH OTHER ORGANIZATIONS

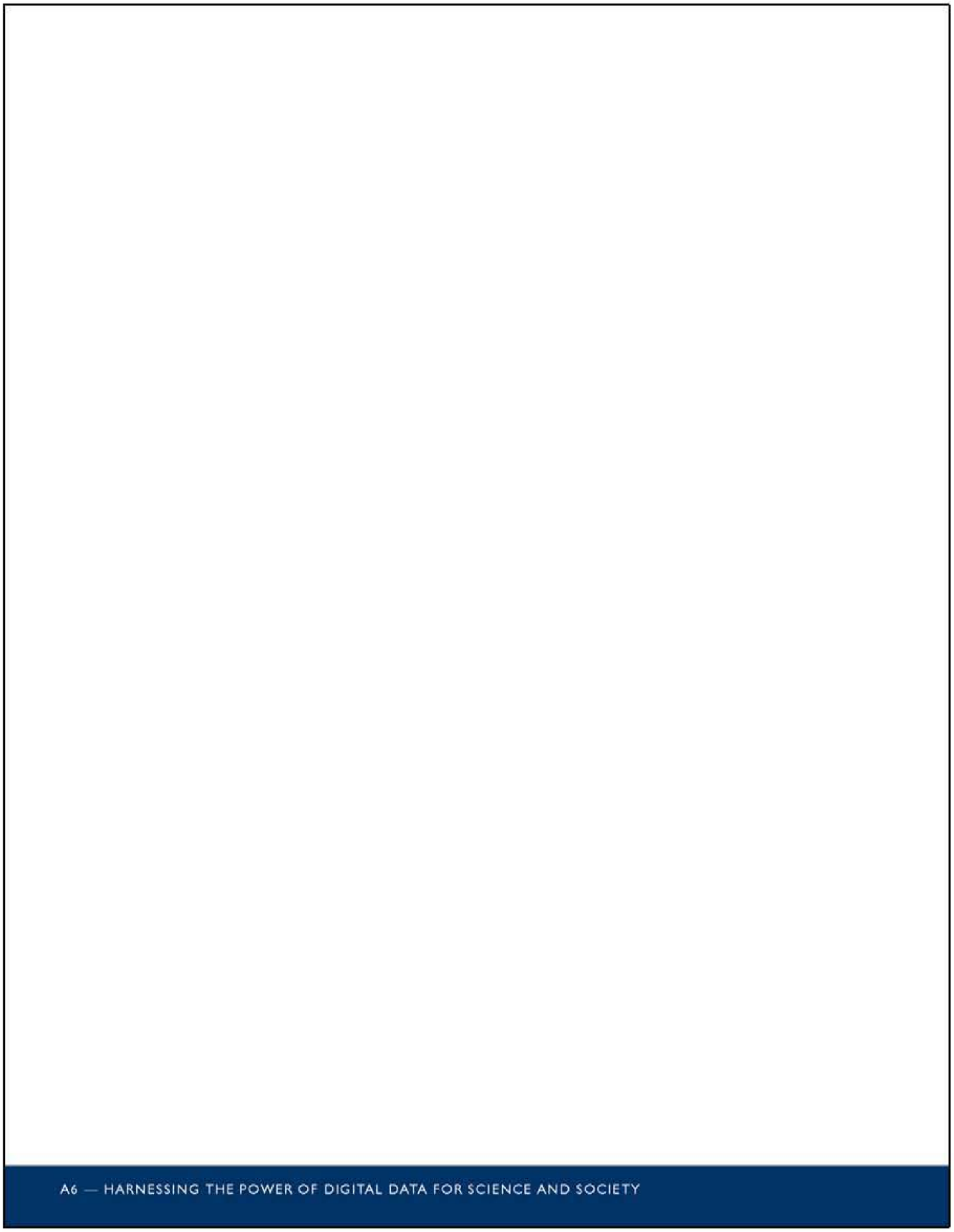
The IWG may interact with other government organizations including the NSTC Committee on Technology (COT), the Networking and Information Technology R&D (NITRD) Subcommittee, which reports to the COT, and the NITRD National Coordination Office. The IWG may also interact with federal advisory bodies such as the President's Council of Advisors on Science and Technology (PCAST). The IWG may interact with and receive ad hoc advice from other interagency groups such as CENDI and the Federal CIO Council, and from private sector groups, professional societies, and other non-government organizations such as the National Academies of Science and Engineering, the Institute of Medicine, and the National Research Council as consistent with the *Federal Advisory Committee Act.*

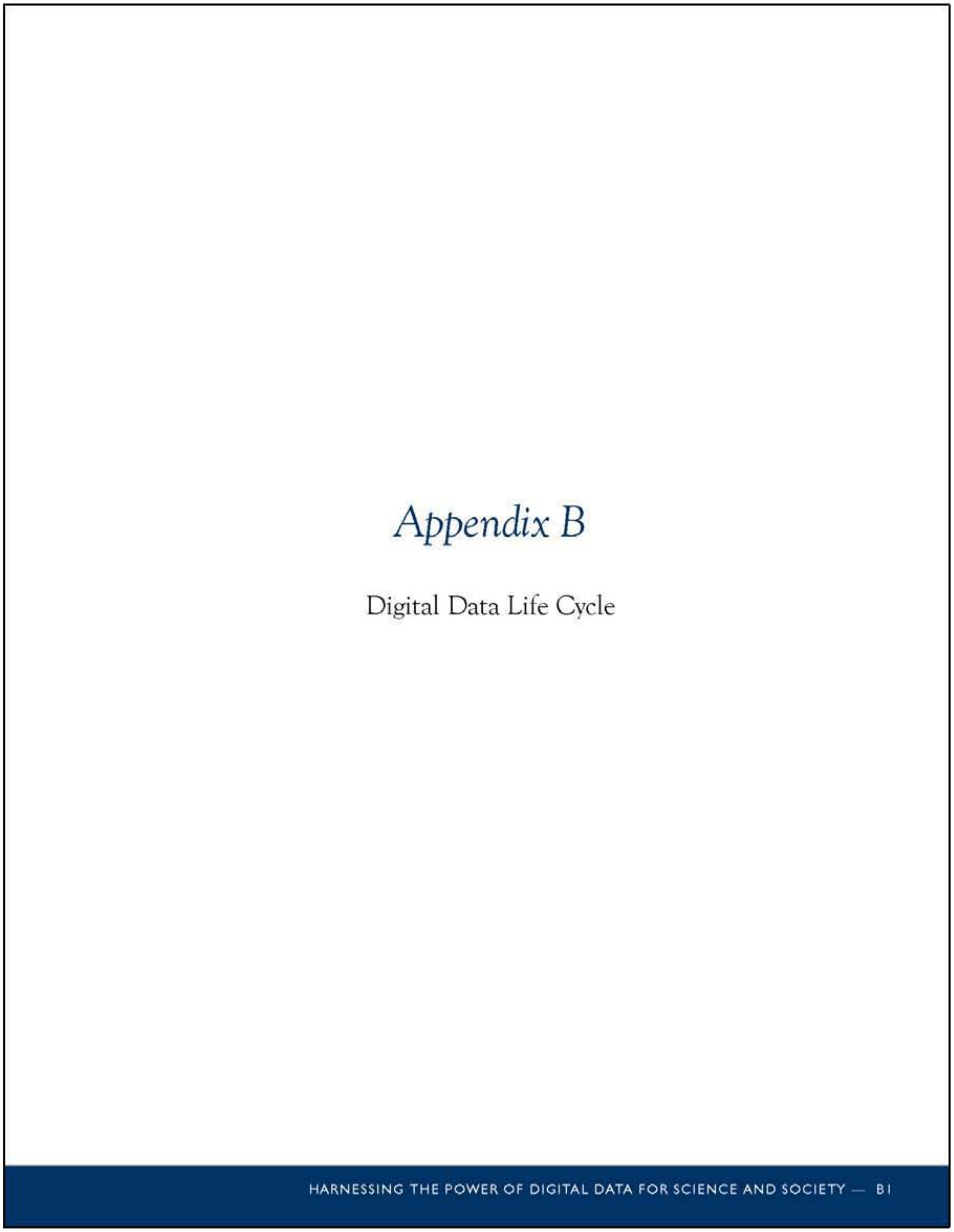
TERMINATION

Unless renewed by the Co-Chairs of the Committee on Science prior to its expiration, the IWG shall terminate no later than March 31, 2009.

DETERMINATION

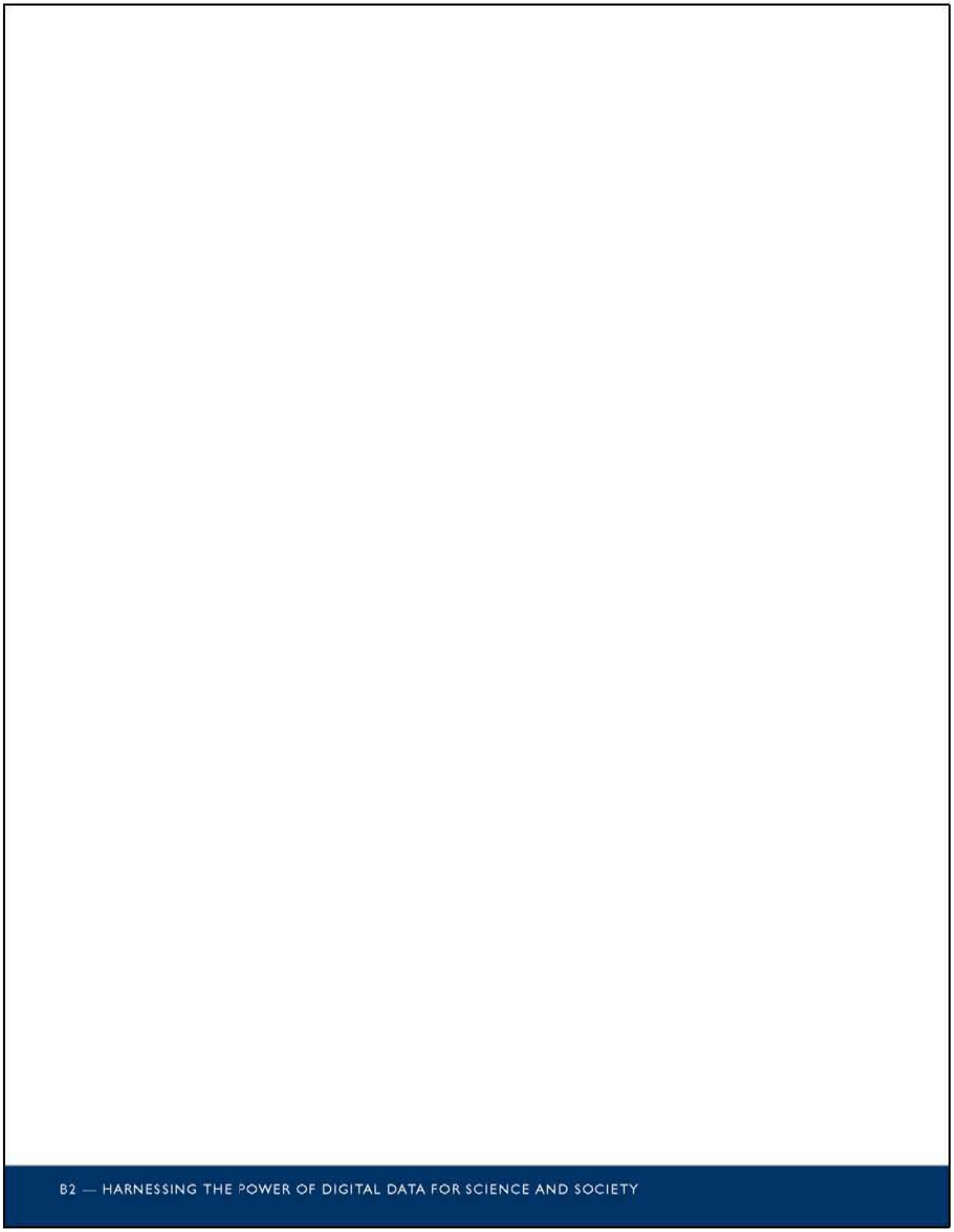
We hereby determine that the formation of this Interagency Working Group is in the public interest in connection with the performance of duties imposed on the Executive Branch by law, and that such duties can best be performed by such a group.



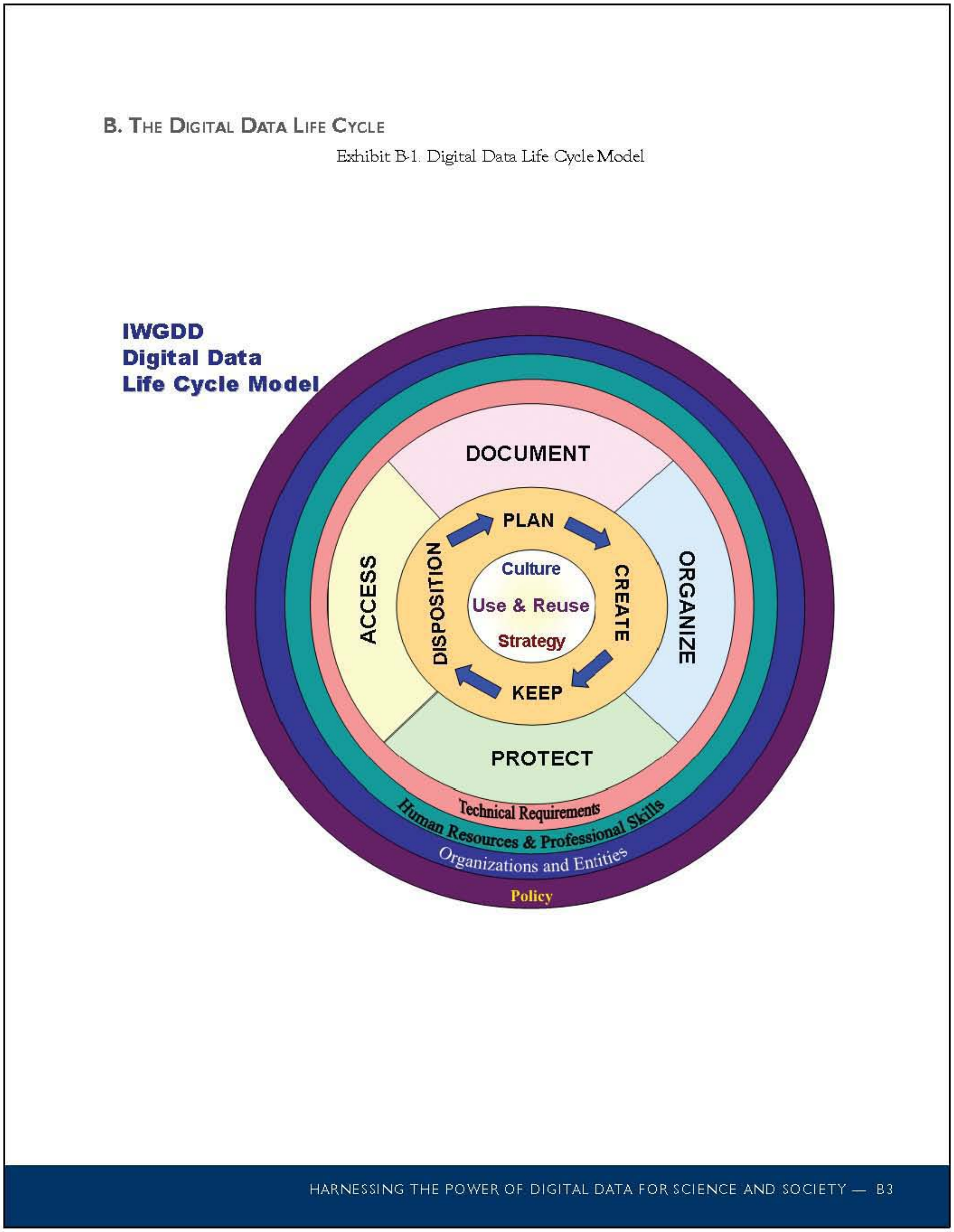


Appendix B

Digital Data Life Cycle



B2 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



B. The Digital Data Life Cycle

Exhibit B-1. Digital Data Life Cycle Model

HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY — B3

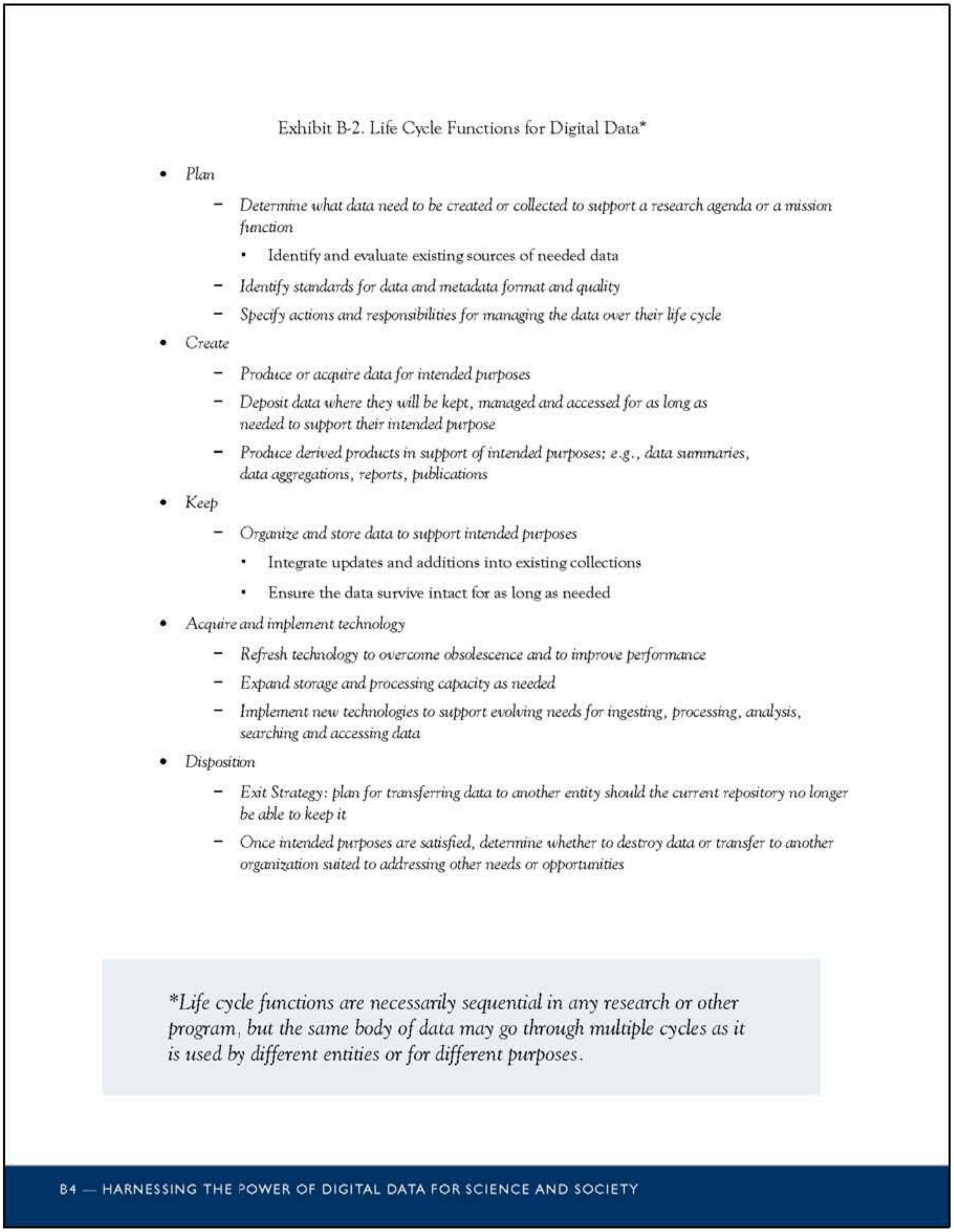


Exhibit B-2. Life Cycle Functions for Digital Data*

- *Plan*
  - *Determine what data need to be created or collected to support a research agenda or a mission function*
    - Identify and evaluate existing sources of needed data
  - *Identify standards for data and metadata format and quality*
  - *Specify actions and responsibilities for managing the data over their life cycle*
- *Create*
  - *Produce or acquire data for intended purposes*
  - *Deposit data where they will be kept, managed and accessed for as long as needed to support their intended purpose*
  - *Produce derived products in support of intended purposes; e.g., data summaries, data aggregations, reports, publications*
- *Keep*
  - *Organize and store data to support intended purposes*
    - Integrate updates and additions into existing collections
    - Ensure the data survive intact for as long as needed
- *Acquire and implement technology*
  - *Refresh technology to overcome obsolescence and to improve performance*
  - *Expand storage and processing capacity as needed*
  - *Implement new technologies to support evolving needs for ingesting, processing, analysis, searching and accessing data*
- *Disposition*
  - *Exit Strategy: plan for transferring data to another entity should the current repository no longer be able to keep it*
  - *Once intended purposes are satisfied, determine whether to destroy data or transfer to another organization suited to addressing other needs or opportunities*

**Life cycle functions are necessarily sequential in any research or other program, but the same body of data may go through multiple cycles as it is used by different entities or for different purposes.*

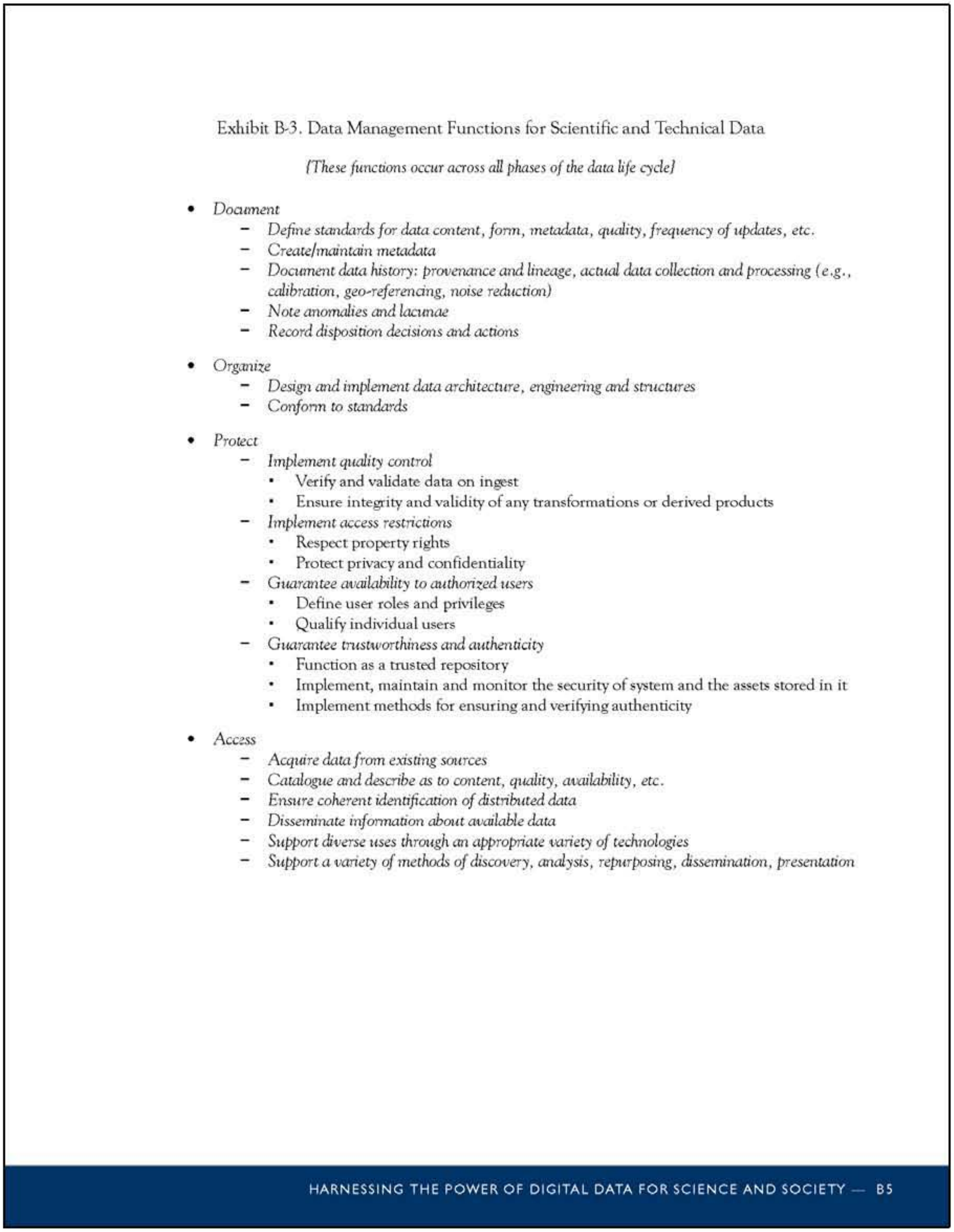
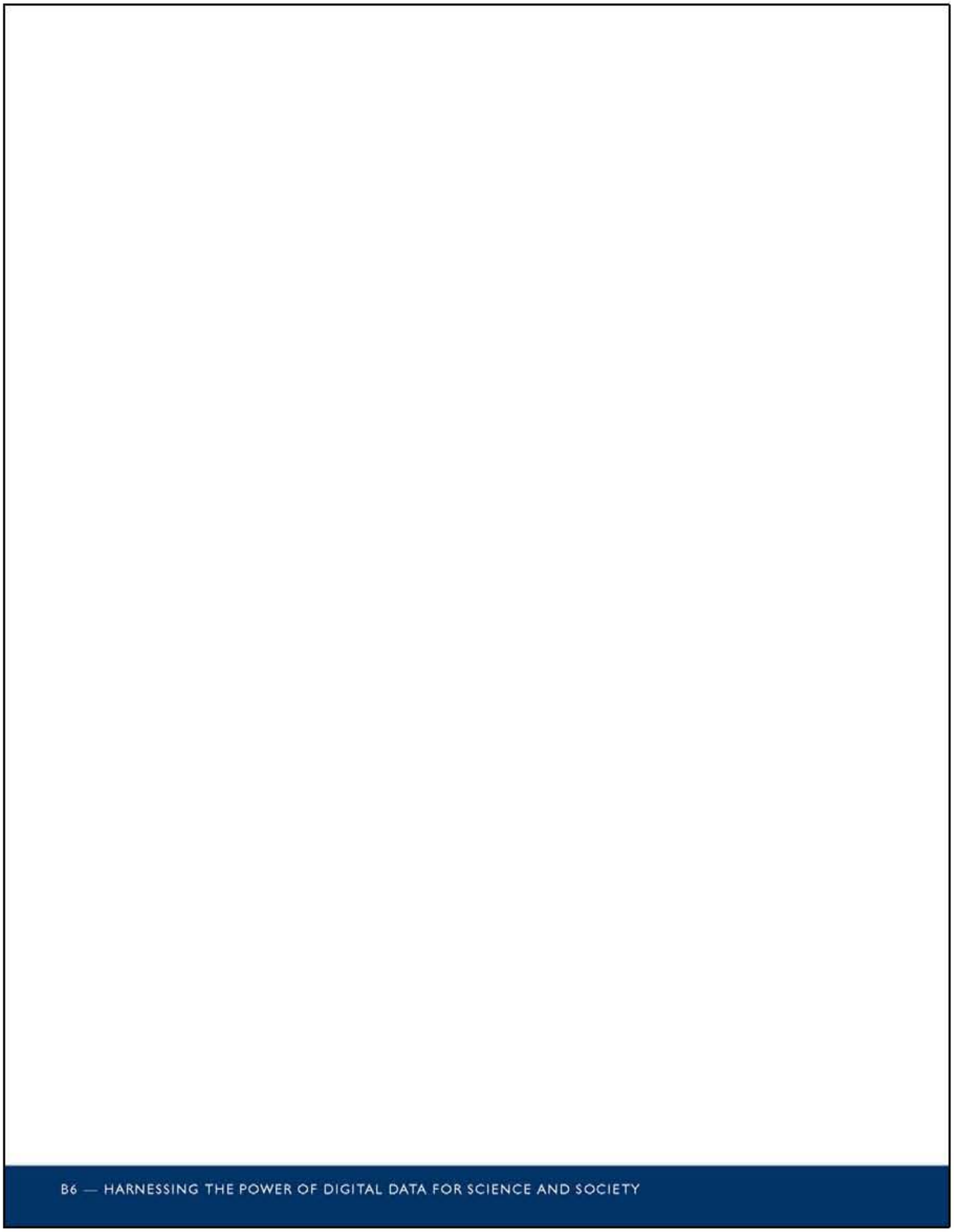


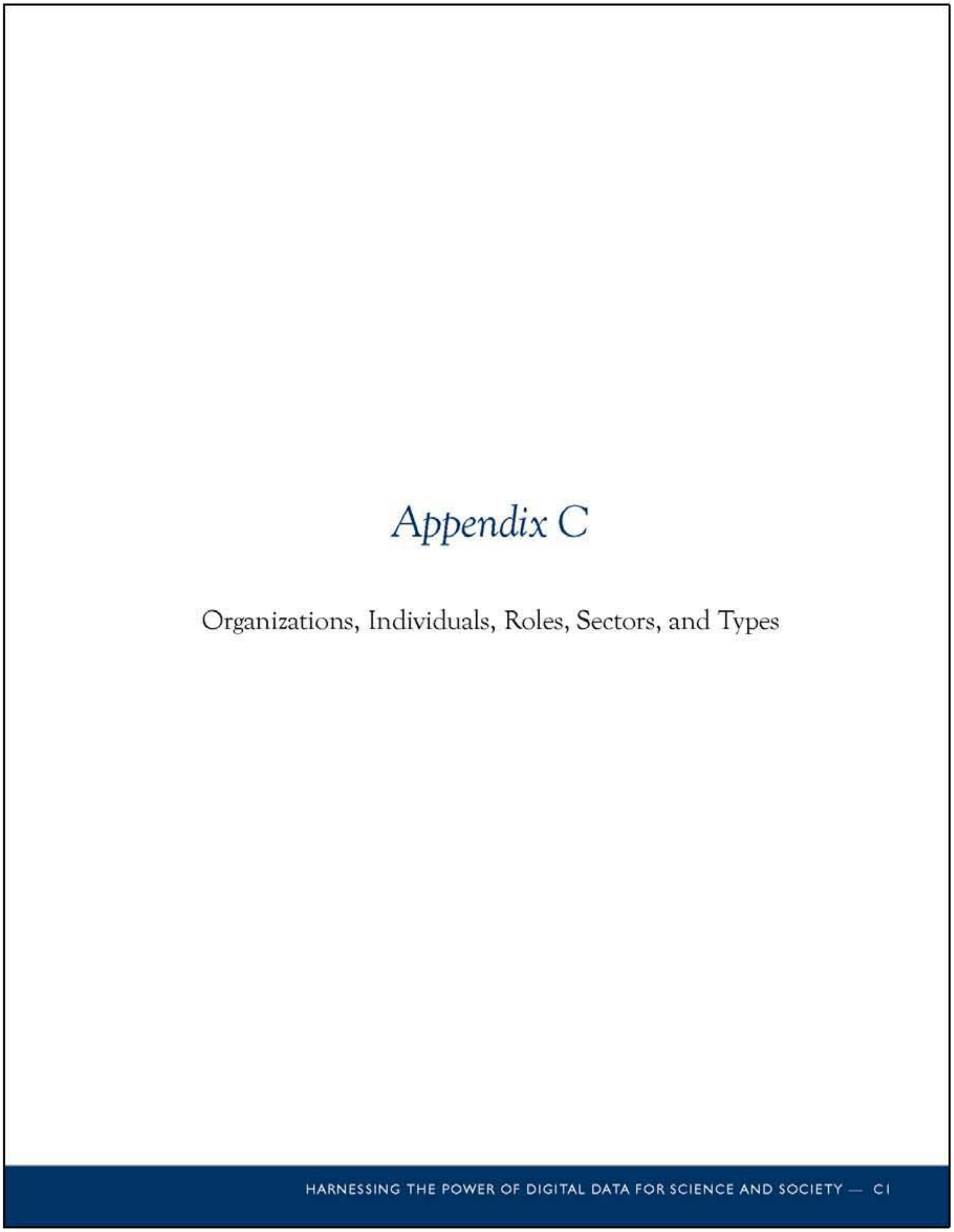
Exhibit B-3. Data Management Functions for Scientific and Technical Data

[These functions occur across all phases of the data life cycle]

- *Document*
  - *Define standards for data content, form, metadata, quality, frequency of updates, etc.*
  - *Create/maintain metadata*
  - *Document data history: provenance and lineage, actual data collection and processing (e.g., calibration, geo-referencing, noise reduction)*
  - *Note anomalies and lacunae*
  - *Record disposition decisions and actions*
- *Organize*
  - *Design and implement data architecture, engineering and structures*
  - *Conform to standards*
- *Protect*
  - *Implement quality control*
    - Verify and validate data on ingest
    - Ensure integrity and validity of any transformations or derived products
  - *Implement access restrictions*
    - Respect property rights
    - Protect privacy and confidentiality
  - *Guarantee availability to authorized users*
    - Define user roles and privileges
    - Qualify individual users
  - *Guarantee trustworthiness and authenticity*
    - Function as a trusted repository
    - Implement, maintain and monitor the security of system and the assets stored in it
    - Implement methods for ensuring and verifying authenticity
- *Access*
  - *Acquire data from existing sources*
  - *Catalogue and describe as to content, quality, availability, etc.*
  - *Ensure coherent identification of distributed data*
  - *Disseminate information about available data*
  - *Support diverse uses through an appropriate variety of technologies*
  - *Support a variety of methods of discovery, analysis, repurposing, dissemination, presentation*

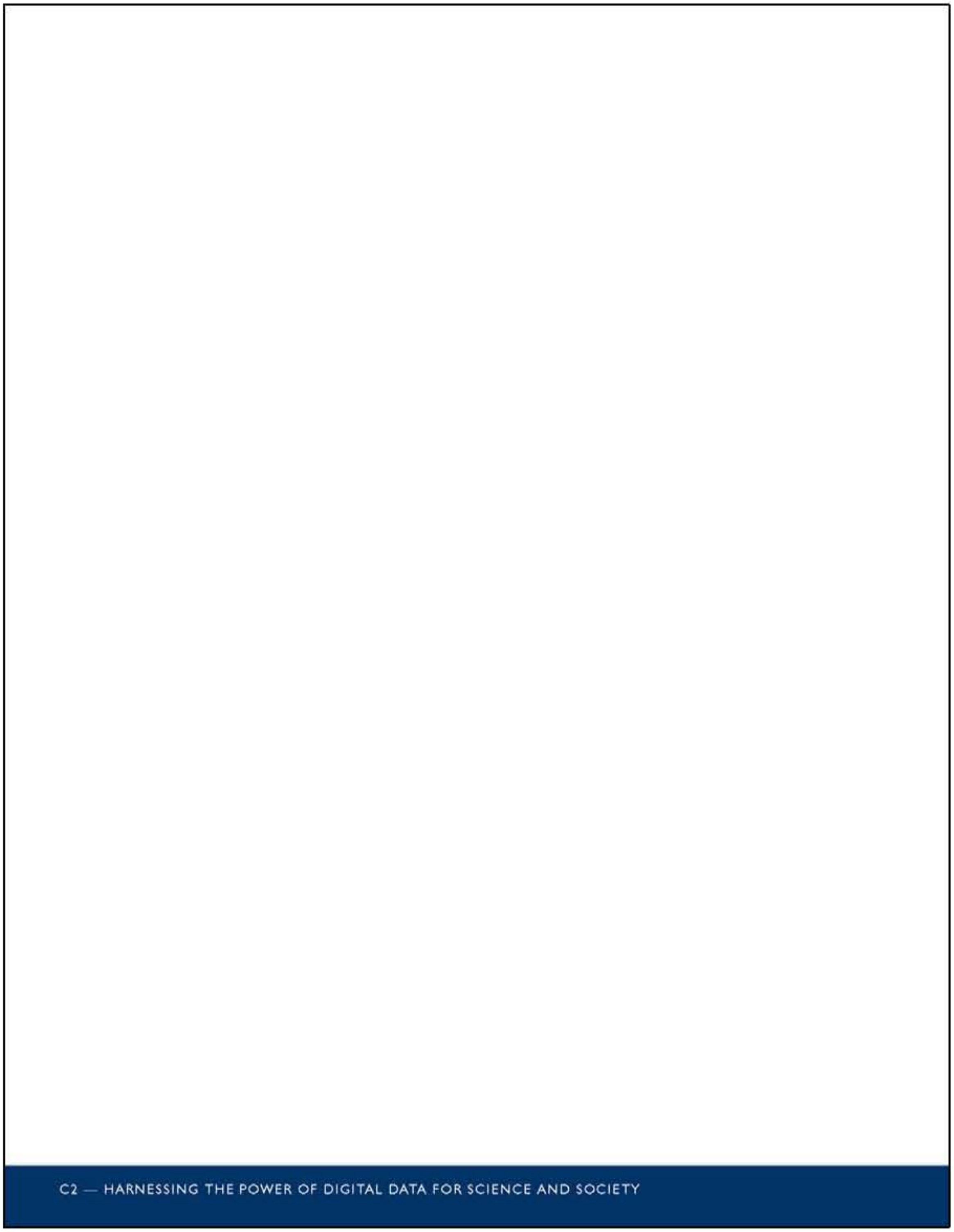


B6 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY

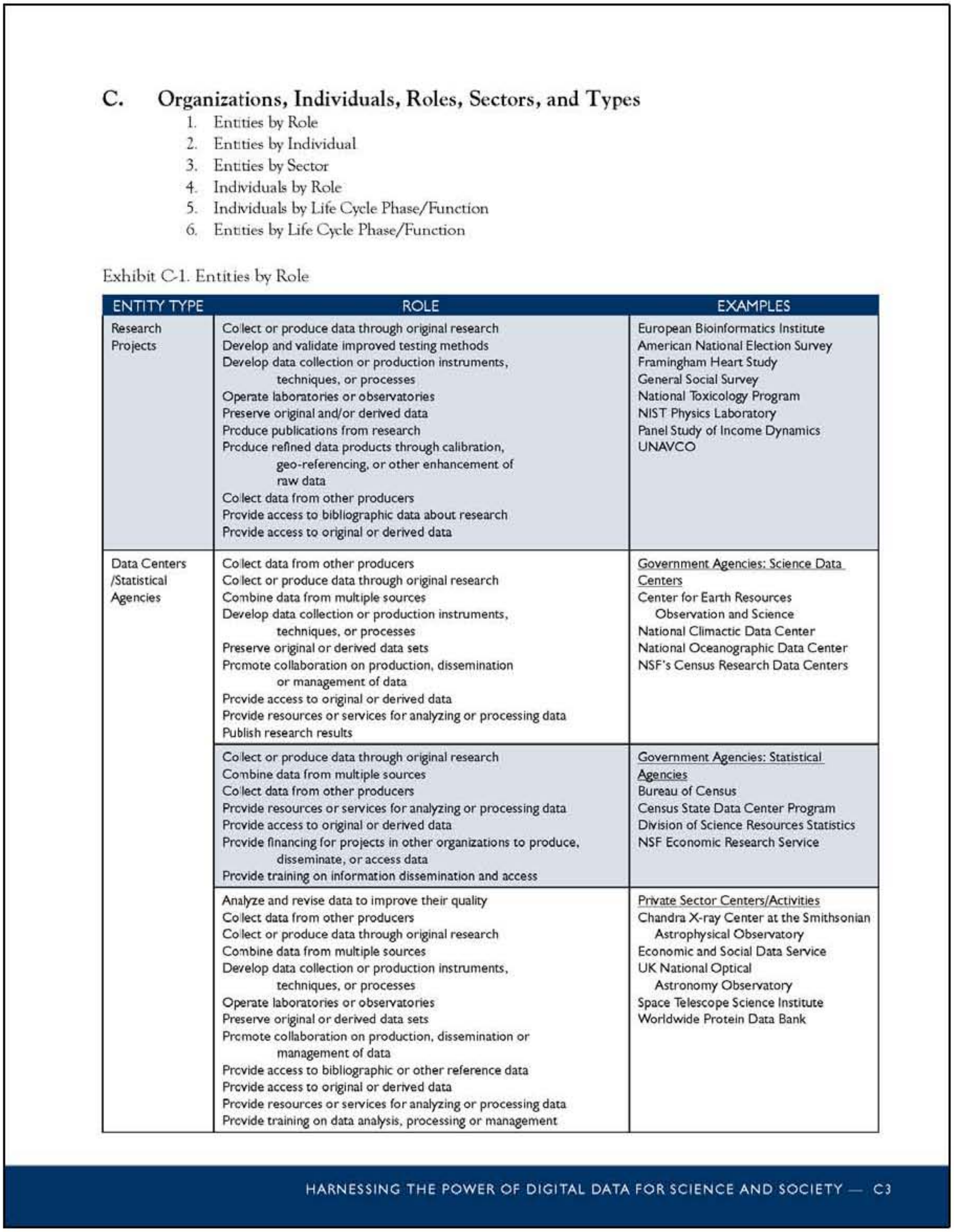


Appendix C

Organizations, Individuals, Roles, Sectors, and Types



C2 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



C. Organizations, Individuals, Roles, Sectors, and Types

1. Entities by Role
2. Entities by Individual
3. Entities by Sector
4. Individuals by Role
5. Individuals by Life Cycle Phase/Function
6. Entities by Life Cycle Phase/Function

Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
Research Projects	Collect or produce data through original research Develop and validate improved testing methods Develop data collection or production instruments, techniques, or processes Operate laboratories or observatories Preserve original and/or derived data Produce publications from research Produce refined data products through calibration, geo-referencing, or other enhancement of raw data Collect data from other producers Provide access to bibliographic data about research Provide access to original or derived data	European Bioinformatics Institute American National Election Survey Framingham Heart Study General Social Survey National Toxicology Program NIST Physics Laboratory Panel Study of Income Dynamics UNAVCO
Data Centers /Statistical Agencies	Collect data from other producers Collect or produce data through original research Combine data from multiple sources Develop data collection or production instruments, techniques, or processes Preserve original or derived data sets Promote collaboration on production, dissemination or management of data Provide access to original or derived data Provide resources or services for analyzing or processing data Publish research results	Government Agencies: Science Data Centers Center for Earth Resources Observation and Science National Climactic Data Center National Oceanographic Data Center NSF's Census Research Data Centers
	Collect or produce data through original research Combine data from multiple sources Collect data from other producers Provide resources or services for analyzing or processing data Provide access to original or derived data Provide financing for projects in other organizations to produce, disseminate, or access data Provide training on information dissemination and access	Government Agencies: Statistical Agencies Bureau of Census Census State Data Center Program Division of Science Resources Statistics NSF Economic Research Service
	Analyze and revise data to improve their quality Collect data from other producers Collect or produce data through original research Combine data from multiple sources Develop data collection or production instruments, techniques, or processes Operate laboratories or observatories Preserve original or derived data sets Promote collaboration on production, dissemination or management of data Provide access to bibliographic or other reference data Provide access to original or derived data Provide resources or services for analyzing or processing data Provide training on data analysis, processing or management	Private Sector Centers/Activities Chandra X-ray Center at the Smithsonian Astrophysical Observatory Economic and Social Data Service UK National Optical Astronomy Observatory Space Telescope Science Institute Worldwide Protein Data Bank

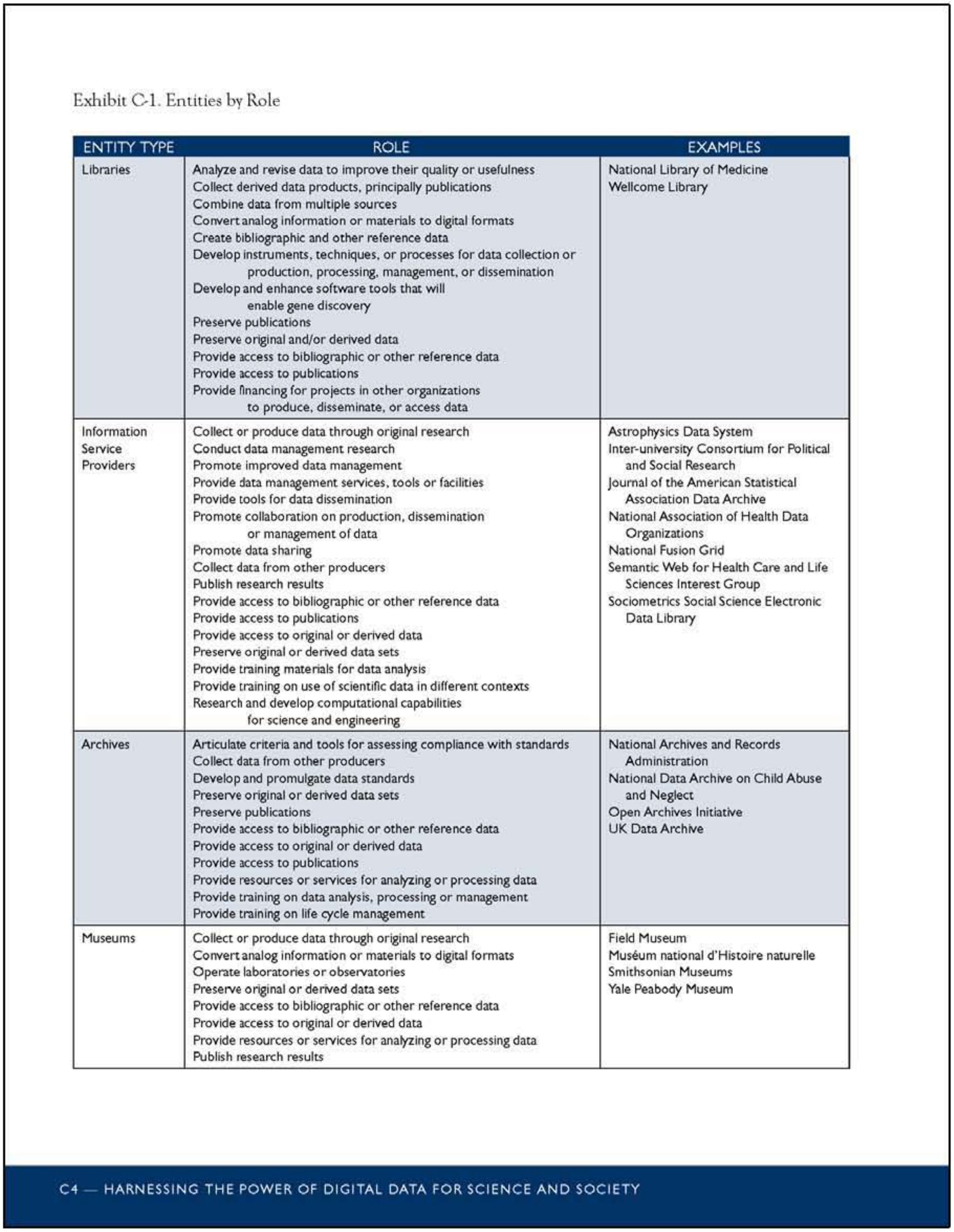


Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
Libraries	Analyze and revise data to improve their quality or usefulness Collect derived data products, principally publications Combine data from multiple sources Convert analog information or materials to digital formats Create bibliographic and other reference data Develop instruments, techniques, or processes for data collection or production, processing, management, or dissemination Develop and enhance software tools that will enable gene discovery Preserve publications Preserve original and/or derived data Provide access to bibliographic or other reference data Provide access to publications Provide financing for projects in other organizations to produce, disseminate, or access data	National Library of Medicine Wellcome Library
Information Service Providers	Collect or produce data through original research Conduct data management research Promote improved data management Provide data management services, tools or facilities Provide tools for data dissemination Promote collaboration on production, dissemination or management of data Promote data sharing Collect data from other producers Publish research results Provide access to bibliographic or other reference data Provide access to publications Provide access to original or derived data Preserve original or derived data sets Provide training materials for data analysis Provide training on use of scientific data in different contexts Research and develop computational capabilities for science and engineering	Astrophysics Data System Inter-university Consortium for Political and Social Research Journal of the American Statistical Association Data Archive National Association of Health Data Organizations National Fusion Grid Semantic Web for Health Care and Life Sciences Interest Group Sociometrics Social Science Electronic Data Library
Archives	Articulate criteria and tools for assessing compliance with standards Collect data from other producers Develop and promulgate data standards Preserve original or derived data sets Preserve publications Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications Provide resources or services for analyzing or processing data Provide training on data analysis, processing or management Provide training on life cycle management	National Archives and Records Administration National Data Archive on Child Abuse and Neglect Open Archives Initiative UK Data Archive
Museums	Collect or produce data through original research Convert analog information or materials to digital formats Operate laboratories or observatories Preserve original or derived data sets Provide access to bibliographic or other reference data Provide access to original or derived data Provide resources or services for analyzing or processing data Publish research results	Field Museum Muséum national d'Histoire naturelle Smithsonian Museums Yale Peabody Museum

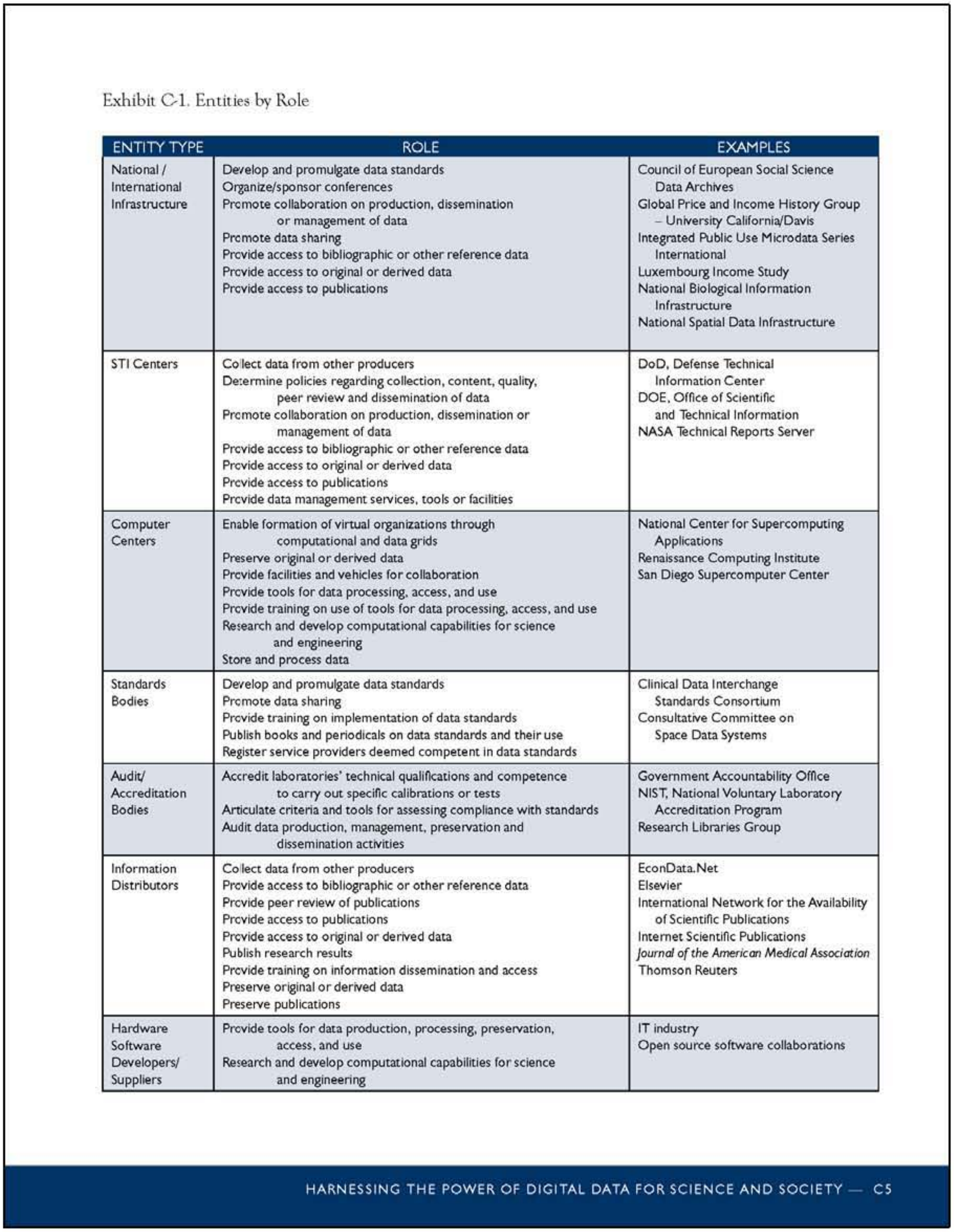


Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
National / International Infrastructure	Develop and promulgate data standards Organize/sponsor conferences Promote collaboration on production, dissemination or management of data Promote data sharing Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications	Council of European Social Science Data Archives Global Price and Income History Group – University California/Davis Integrated Public Use Microdata Series International Luxembourg Income Study National Biological Information Infrastructure National Spatial Data Infrastructure
STI Centers	Collect data from other producers Determine policies regarding collection, content, quality, peer review and dissemination of data Promote collaboration on production, dissemination or management of data Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications Provide data management services, tools or facilities	DoD, Defense Technical Information Center DOE, Office of Scientific and Technical Information NASA Technical Reports Server
Computer Centers	Enable formation of virtual organizations through computational and data grids Preserve original or derived data Provide facilities and vehicles for collaboration Provide tools for data processing, access, and use Provide training on use of tools for data processing, access, and use Research and develop computational capabilities for science and engineering Store and process data	National Center for Supercomputing Applications Renaissance Computing Institute San Diego Supercomputer Center
Standards Bodies	Develop and promulgate data standards Promote data sharing Provide training on implementation of data standards Publish books and periodicals on data standards and their use Register service providers deemed competent in data standards	Clinical Data Interchange Standards Consortium Consultative Committee on Space Data Systems
Audit/ Accreditation Bodies	Accredit laboratories' technical qualifications and competence to carry out specific calibrations or tests Articulate criteria and tools for assessing compliance with standards Audit data production, management, preservation and dissemination activities	Government Accountability Office NIST, National Voluntary Laboratory Accreditation Program Research Libraries Group
Information Distributors	Collect data from other producers Provide access to bibliographic or other reference data Provide peer review of publications Provide access to publications Provide access to original or derived data Publish research results Provide training on information dissemination and access Preserve original or derived data Preserve publications	EconData.Net Elsevier International Network for the Availability of Scientific Publications Internet Scientific Publications *Journal of the American Medical Association* Thomson Reuters
Hardware Software Developers/ Suppliers	Provide tools for data production, processing, preservation, access, and use Research and develop computational capabilities for science and engineering	IT industry Open source software collaborations

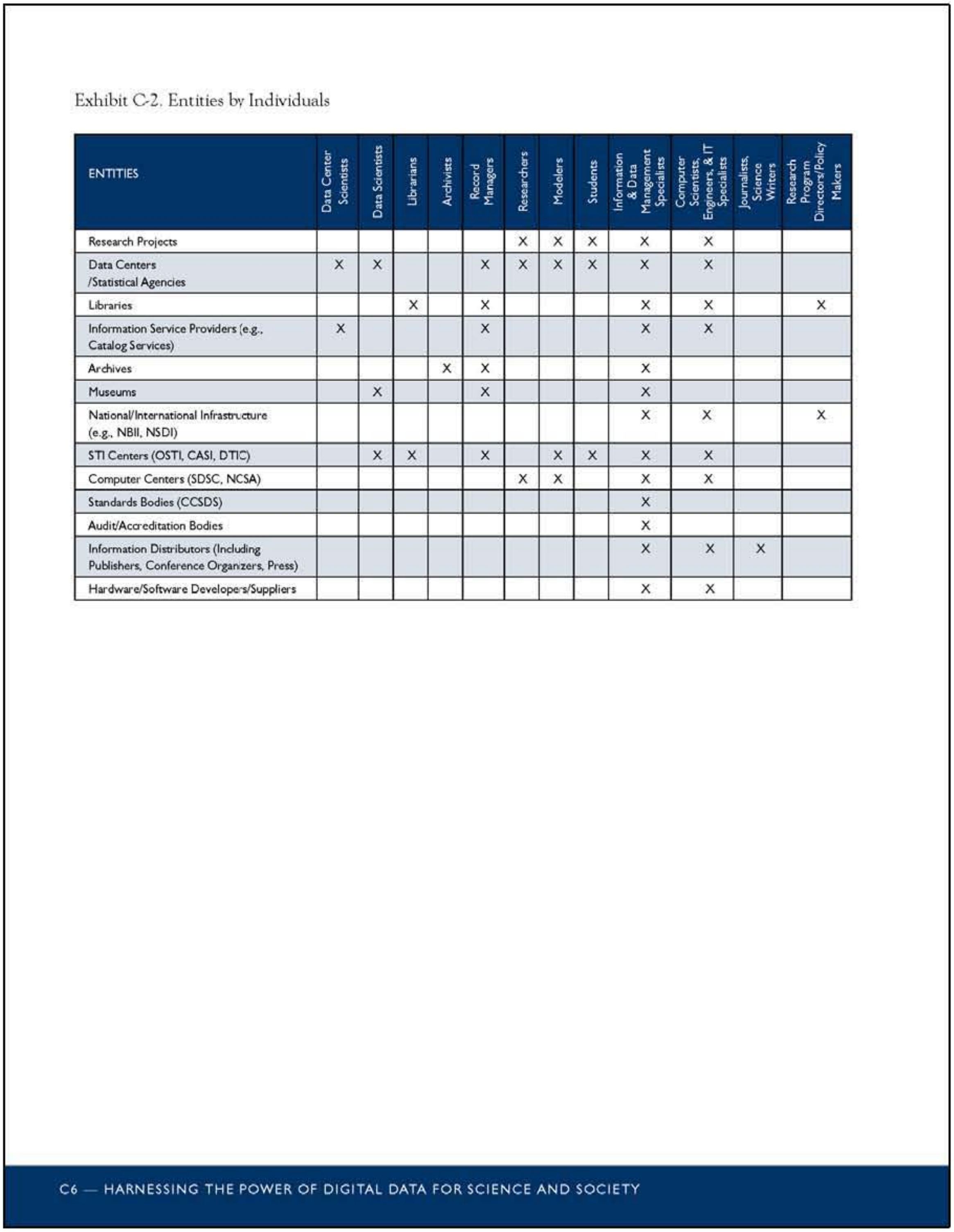


Exhibit C-2. Entities by Individuals

ENTITIES	Data Center Scientists	Data Scientists	Librarians	Archivists	Record Managers	Researchers	Modelers	Students	Information & Data Management Specialists	Computer Scientists, Engineers, & IT Specialists	Journalists, Science Writers	Research Program Directors/Policy Makers
Research Projects						X	X	X	X	X		
Data Centers /Statistical Agencies	X	X			X	X	X	X	X	X		
Libraries			X		X				X	X		X
Information Service Providers (e.g., Catalog Services)	X				X				X	X		
Archives				X	X				X			
Museums		X			X				X			
National/International Infrastructure (e.g., NBII, NSDI)									X	X		X
STI Centers (OSTI, CASI, DTIC)		X	X		X		X	X	X	X		
Computer Centers (SDSC, NCSA)						X	X		X	X		
Standards Bodies (CCSDS)									X			
Audit/Accreditation Bodies									X			
Information Distributors (Including Publishers, Conference Organizers, Press)									X	X	X	
Hardware/Software Developers/Suppliers									X	X		

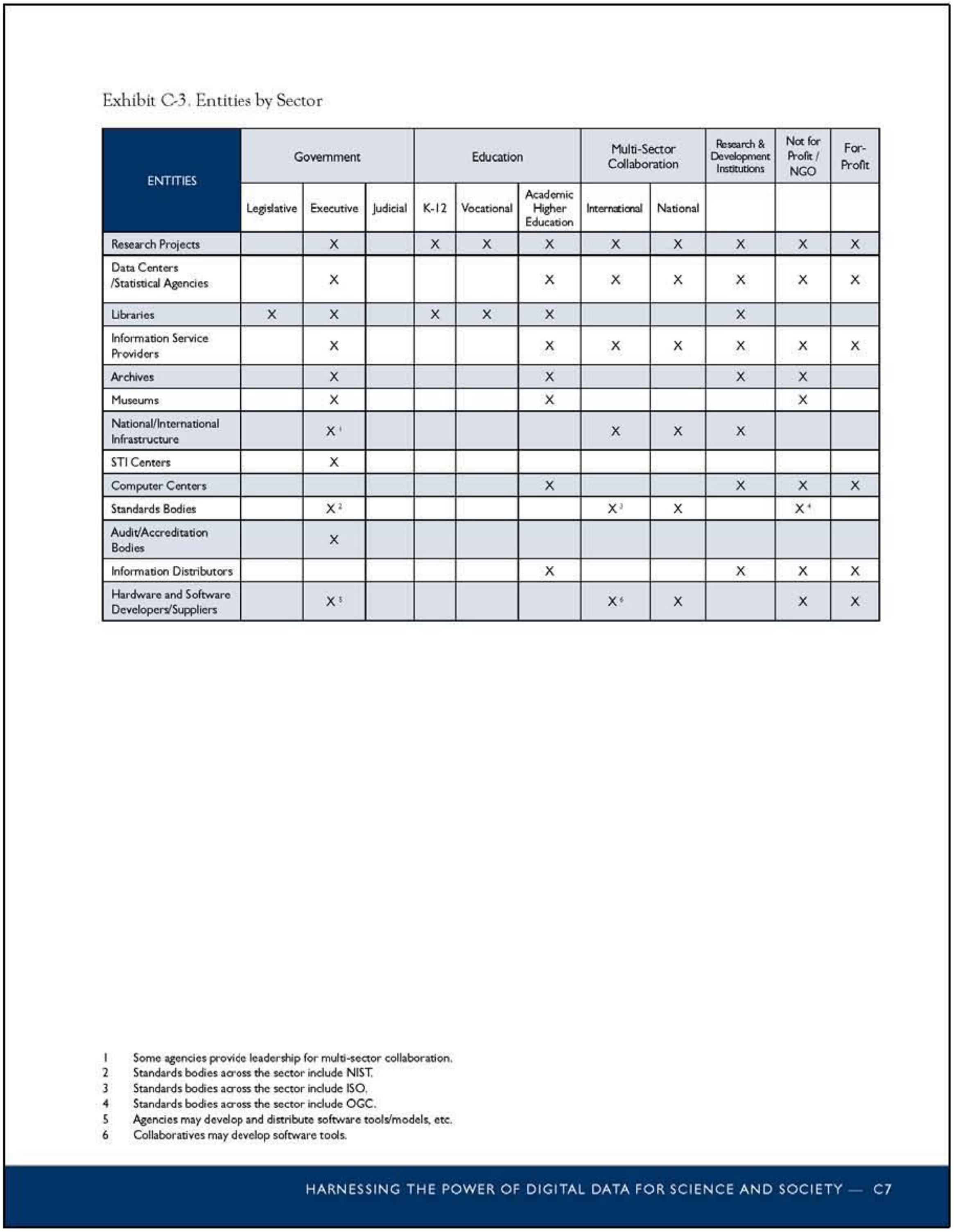


Exhibit C-3. Entities by Sector

ENTITIES	Government			Education			Multi-Sector Collaboration		Research & Development Institutions	Not for Profit / NGO	For-Profit
	Legislative	Executive	Judicial	K-12	Vocational	Academic Higher Education	International	National			
Research Projects		X		X	X	X	X	X	X	X	X
Data Centers /Statistical Agencies		X				X	X	X	X	X	X
Libraries	X	X		X	X	X			X		
Information Service Providers		X				X	X	X	X	X	X
Archives		X				X			X	X	
Museums		X				X				X	
National/International Infrastructure		X[1]					X	X	X		
STI Centers		X									
Computer Centers						X			X	X	X
Standards Bodies		X[2]					X[3]	X		X[4]	
Audit/Accreditation Bodies		X									
Information Distributors						X			X	X	X
Hardware and Software Developers/Suppliers		X[5]					X[6]	X		X	X

[1] Some agencies provide leadership for multi-sector collaboration.
[2] Standards bodies across the sector include NIST.
[3] Standards bodies across the sector include ISO.
[4] Standards bodies across the sector include OGC.
[5] Agencies may develop and distribute software tools/models, etc.
[6] Collaboratives may develop software tools.

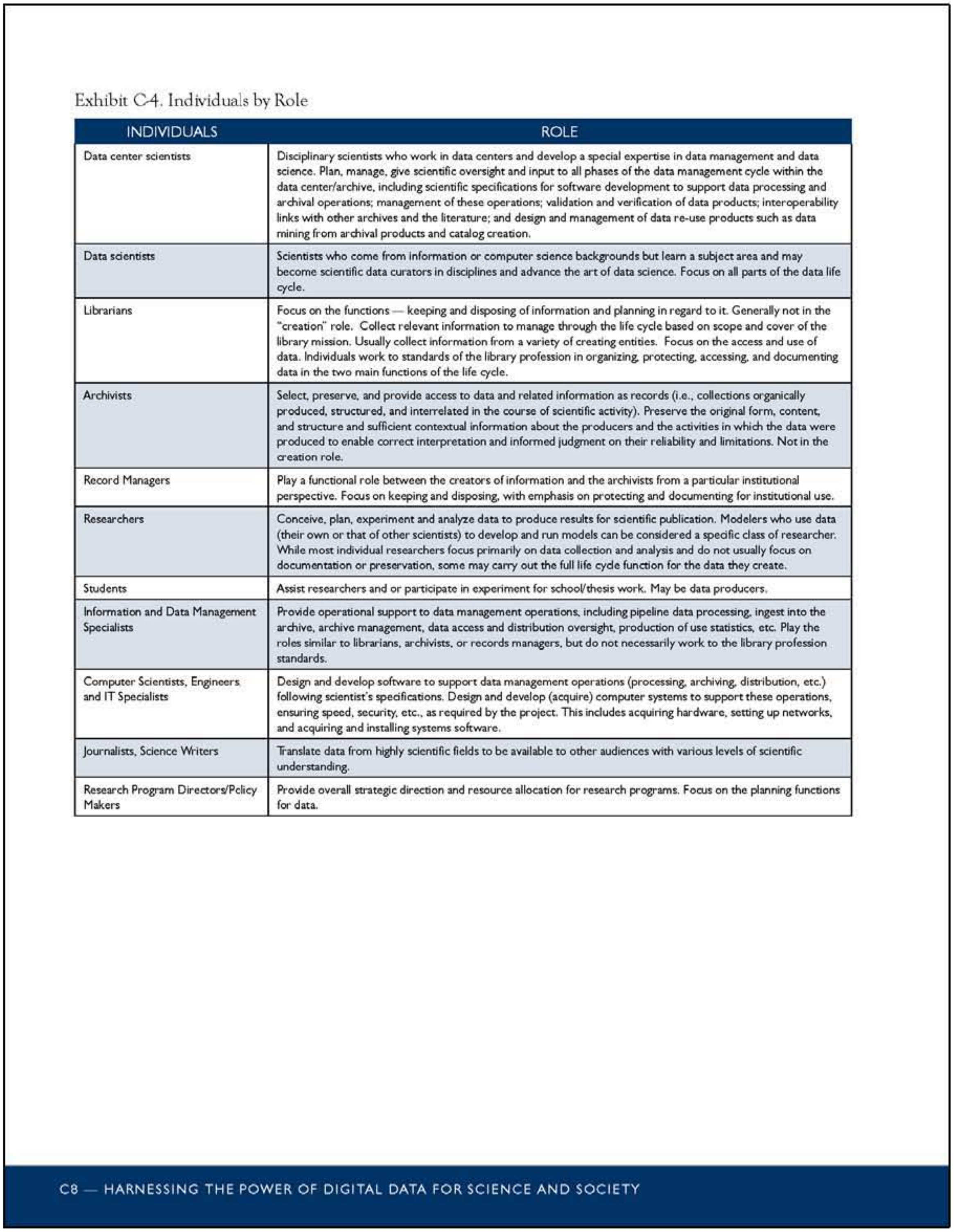


Exhibit C4. Individuals by Role

INDIVIDUALS	ROLE
Data center scientists	Disciplinary scientists who work in data centers and develop a special expertise in data management and data science. Plan, manage, give scientific oversight and input to all phases of the data management cycle within the data center/archive, including scientific specifications for software development to support data processing and archival operations; management of these operations; validation and verification of data products; interoperability links with other archives and the literature; and design and management of data re-use products such as data mining from archival products and catalog creation.
Data scientists	Scientists who come from information or computer science backgrounds but learn a subject area and may become scientific data curators in disciplines and advance the art of data science. Focus on all parts of the data life cycle.
Librarians	Focus on the functions — keeping and disposing of information and planning in regard to it. Generally not in the "creation" role. Collect relevant information to manage through the life cycle based on scope and cover of the library mission. Usually collect information from a variety of creating entities. Focus on the access and use of data. Individuals work to standards of the library profession in organizing, protecting, accessing, and documenting data in the two main functions of the life cycle.
Archivists	Select, preserve, and provide access to data and related information as records (i.e., collections organically produced, structured, and interrelated in the course of scientific activity). Preserve the original form, content, and structure and sufficient contextual information about the producers and the activities in which the data were produced to enable correct interpretation and informed judgment on their reliability and limitations. Not in the creation role.
Record Managers	Play a functional role between the creators of information and the archivists from a particular institutional perspective. Focus on keeping and disposing, with emphasis on protecting and documenting for institutional use.
Researchers	Conceive, plan, experiment and analyze data to produce results for scientific publication. Modelers who use data (their own or that of other scientists) to develop and run models can be considered a specific class of researcher. While most individual researchers focus primarily on data collection and analysis and do not usually focus on documentation or preservation, some may carry out the full life cycle function for the data they create.
Students	Assist researchers and or participate in experiment for school/thesis work. May be data producers.
Information and Data Management Specialists	Provide operational support to data management operations, including pipeline data processing, ingest into the archive, archive management, data access and distribution oversight, production of use statistics, etc. Play the roles similar to librarians, archivists, or records managers, but do not necessarily work to the library profession standards.
Computer Scientists, Engineers and IT Specialists	Design and develop software to support data management operations (processing, archiving, distribution, etc.) following scientist's specifications. Design and develop (acquire) computer systems to support these operations, ensuring speed, security, etc., as required by the project. This includes acquiring hardware, setting up networks, and acquiring and installing systems software.
Journalists, Science Writers	Translate data from highly scientific fields to be available to other audiences with various levels of scientific understanding.
Research Program Directors/Pclicy Makers	Provide overall strategic direction and resource allocation for research programs. Focus on the planning functions for data.

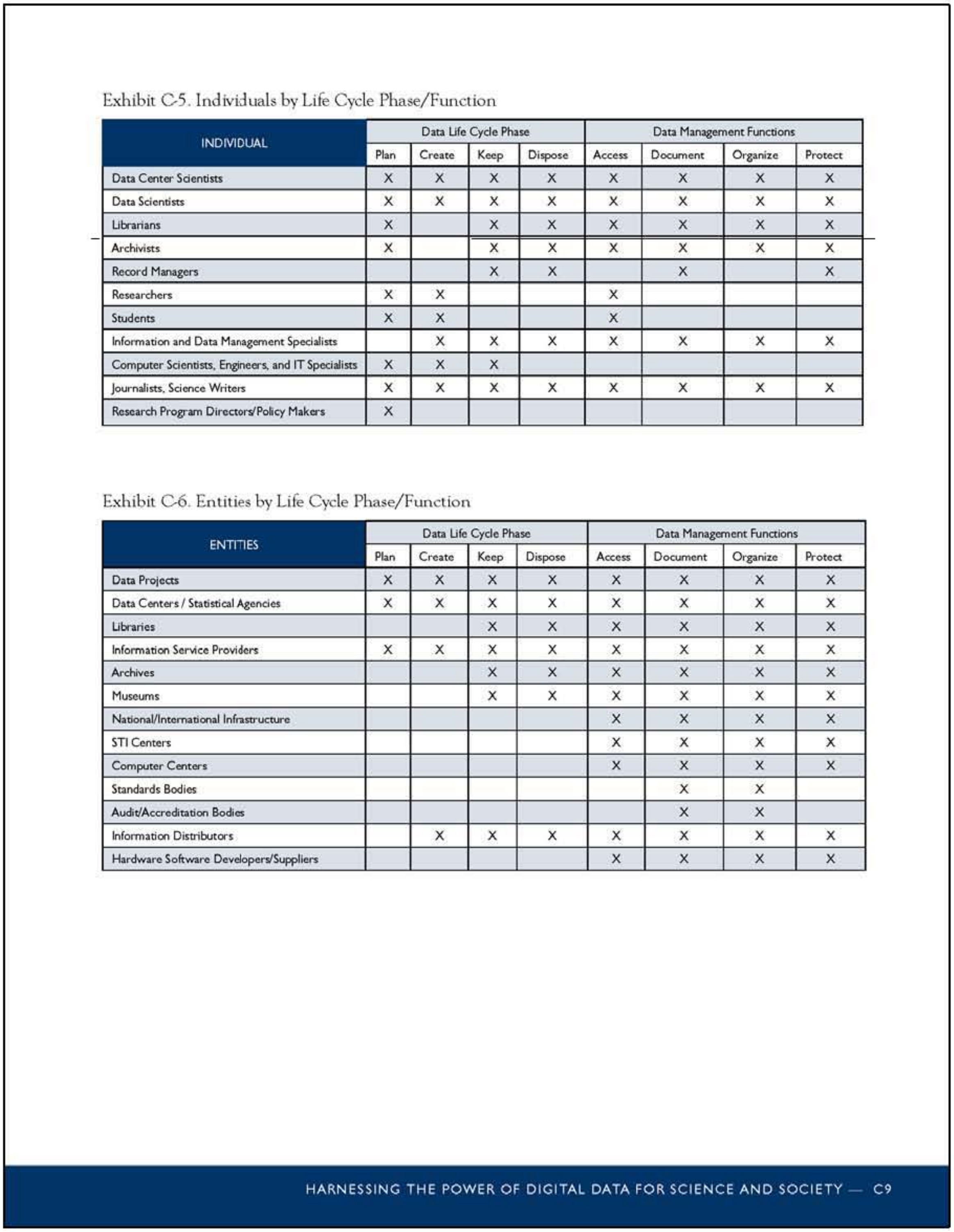
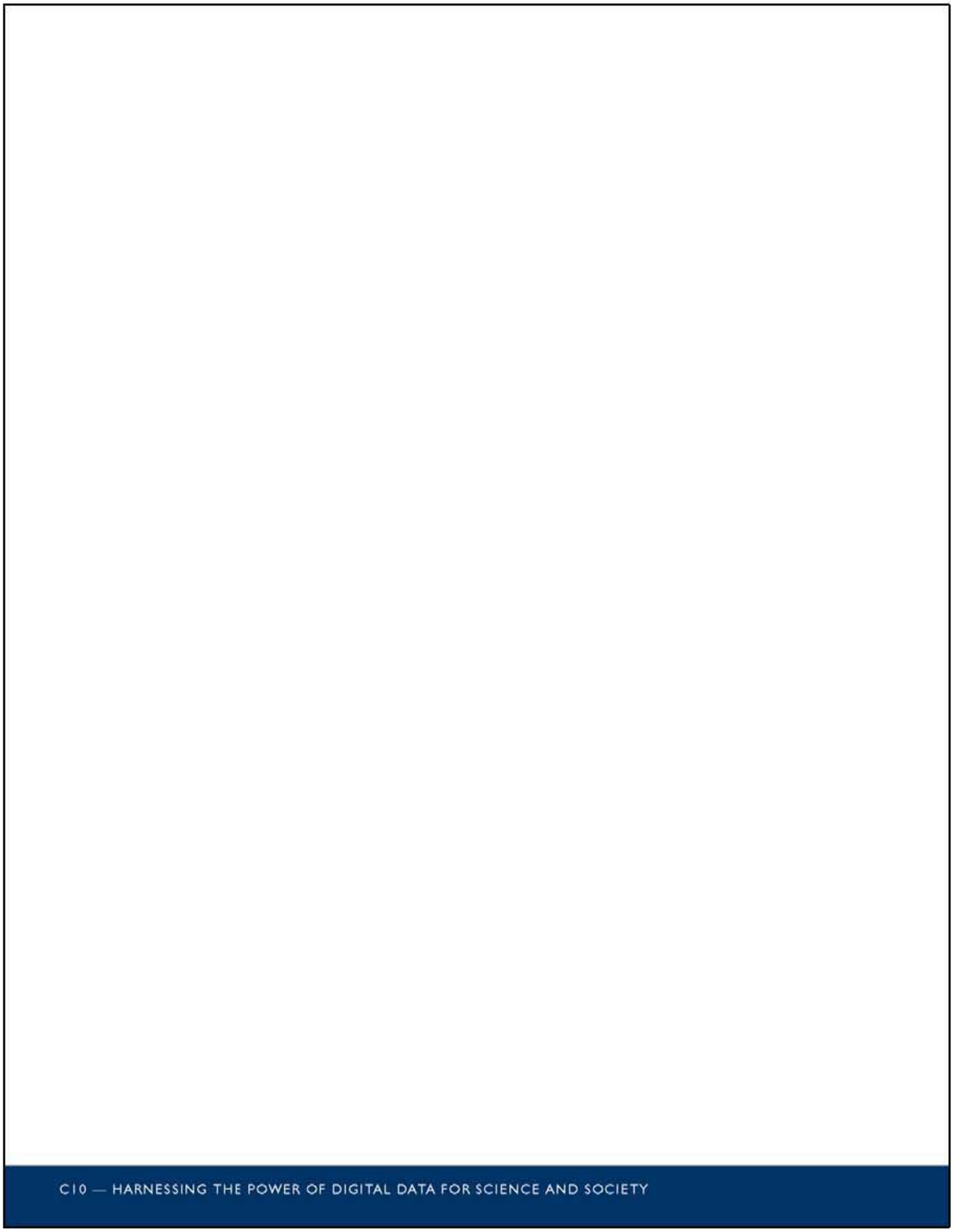


Exhibit C-5. Individuals by Life Cycle Phase/Function

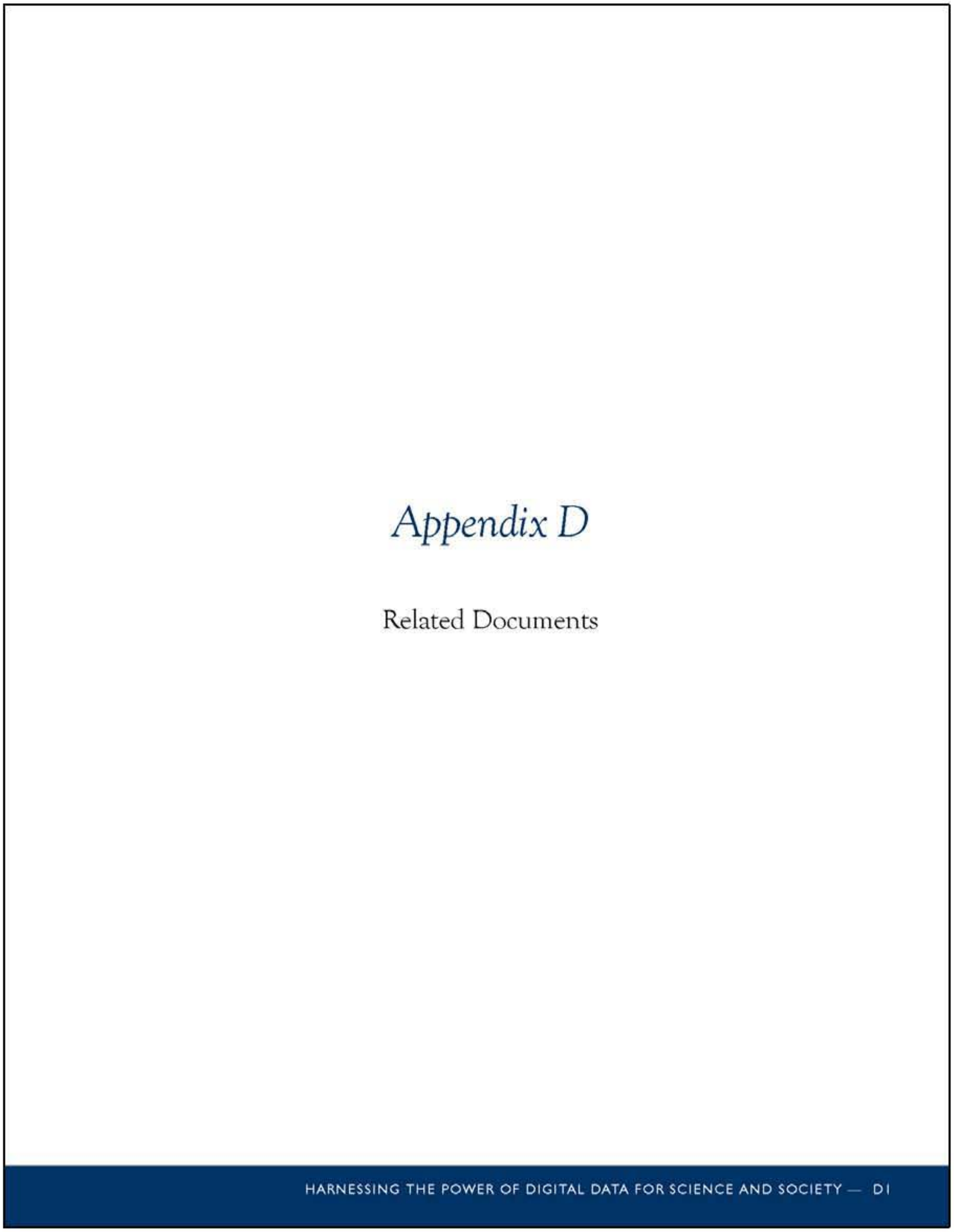
INDIVIDUAL	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Center Scientists	X	X	X	X	X	X	X	X
Data Scientists	X	X	X	X	X	X	X	X
Librarians	X		X	X	X	X	X	X
Archivists	X		X	X	X	X	X	X
Record Managers			X	X		X		X
Researchers	X	X			X			
Students	X	X			X			
Information and Data Management Specialists		X	X	X	X	X	X	X
Computer Scientists, Engineers, and IT Specialists	X	X	X					
Journalists, Science Writers	X	X	X	X	X	X	X	X
Research Program Directors/Policy Makers	X							

Exhibit C-6. Entities by Life Cycle Phase/Function

ENTITIES	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Projects	X	X	X	X	X	X	X	X
Data Centers / Statistical Agencies	X	X	X	X	X	X	X	X
Libraries			X	X	X	X	X	X
Information Service Providers	X	X	X	X	X	X	X	X
Archives			X	X	X	X	X	X
Museums			X	X	X	X	X	X
National/International Infrastructure					X	X	X	X
STI Centers					X	X	X	X
Computer Centers					X	X	X	X
Standards Bodies						X	X	
Audit/Accreditation Bodies						X	X	
Information Distributors		X	X	X	X	X	X	X
Hardware Software Developers/Suppliers					X	X	X	X



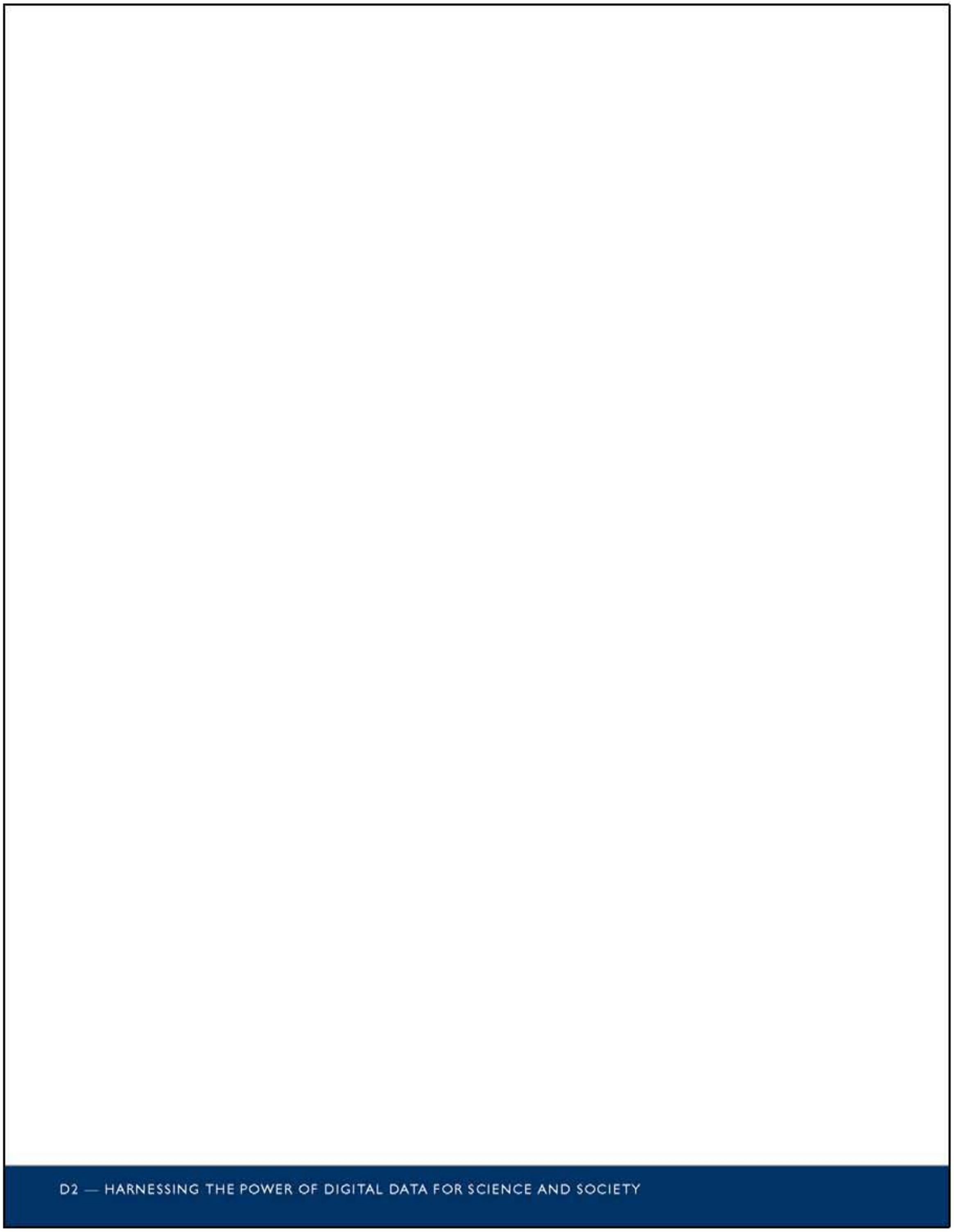
C10 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



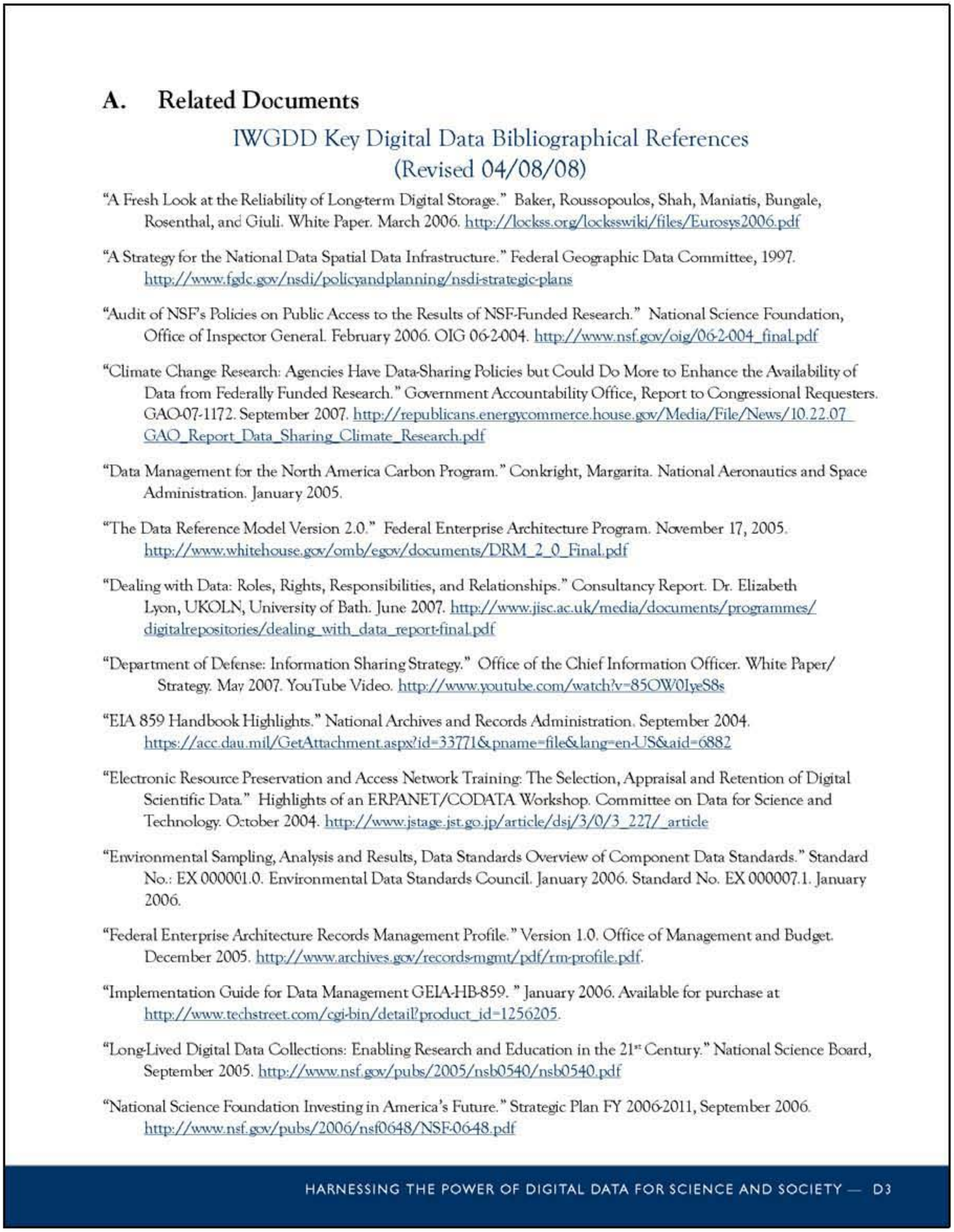
Appendix D

Related Documents

HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY — D1



D2 — HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY



A. Related Documents

IWGDD Key Digital Data Bibliographical References (Revised 04/08/08)

"A Fresh Look at the Reliability of Long-term Digital Storage." Baker, Roussopoulos, Shah, Maniatis, Bungale, Rosenthal, and Giuli. White Paper. March 2006. http://lockss.org/lockswiki/files/Eurosys2006.pdf

"A Strategy for the National Data Spatial Data Infrastructure." Federal Geographic Data Committee, 1997. http://www.fgdc.gov/nsdi/policyandplanning/nsdi-strategic-plans

"Audit of NSF's Policies on Public Access to the Results of NSF-Funded Research." National Science Foundation, Office of Inspector General. February 2006. OIG 06-2-004. http://www.nsf.gov/oig/06-2-004_final.pdf

"Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research." Government Accountability Office, Report to Congressional Requesters. GAO-07-1172. September 2007. http://republicans.energycommerce.house.gov/Media/File/News/10.22.07_GAO_Report_Data_Sharing_Climate_Research.pdf

"Data Management for the North America Carbon Program." Conkright, Margarita. National Aeronautics and Space Administration. January 2005.

"The Data Reference Model Version 2.0." Federal Enterprise Architecture Program. November 17, 2005. http://www.whitehouse.gov/omb/egov/documents/DRM_2_0_Final.pdf

"Dealing with Data: Roles, Rights, Responsibilities, and Relationships." Consultancy Report. Dr. Elizabeth Lyon, UKOLN, University of Bath. June 2007. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

"Department of Defense: Information Sharing Strategy." Office of the Chief Information Officer. White Paper/Strategy. May 2007. YouTube Video. http://www.youtube.com/watch?v=85OW0IyeS8s

"EIA 859 Handbook Highlights." National Archives and Records Administration. September 2004. https://acc.dau.mil/GetAttachment.aspx?id=33771&pname=file&lang=en-US&aid=6882

"Electronic Resource Preservation and Access Network Training: The Selection, Appraisal and Retention of Digital Scientific Data." Highlights of an ERPANET/CODATA Workshop. Committee on Data for Science and Technology. October 2004. http://www.jstage.jst.go.jp/article/dsj/3/0/3_227/_article

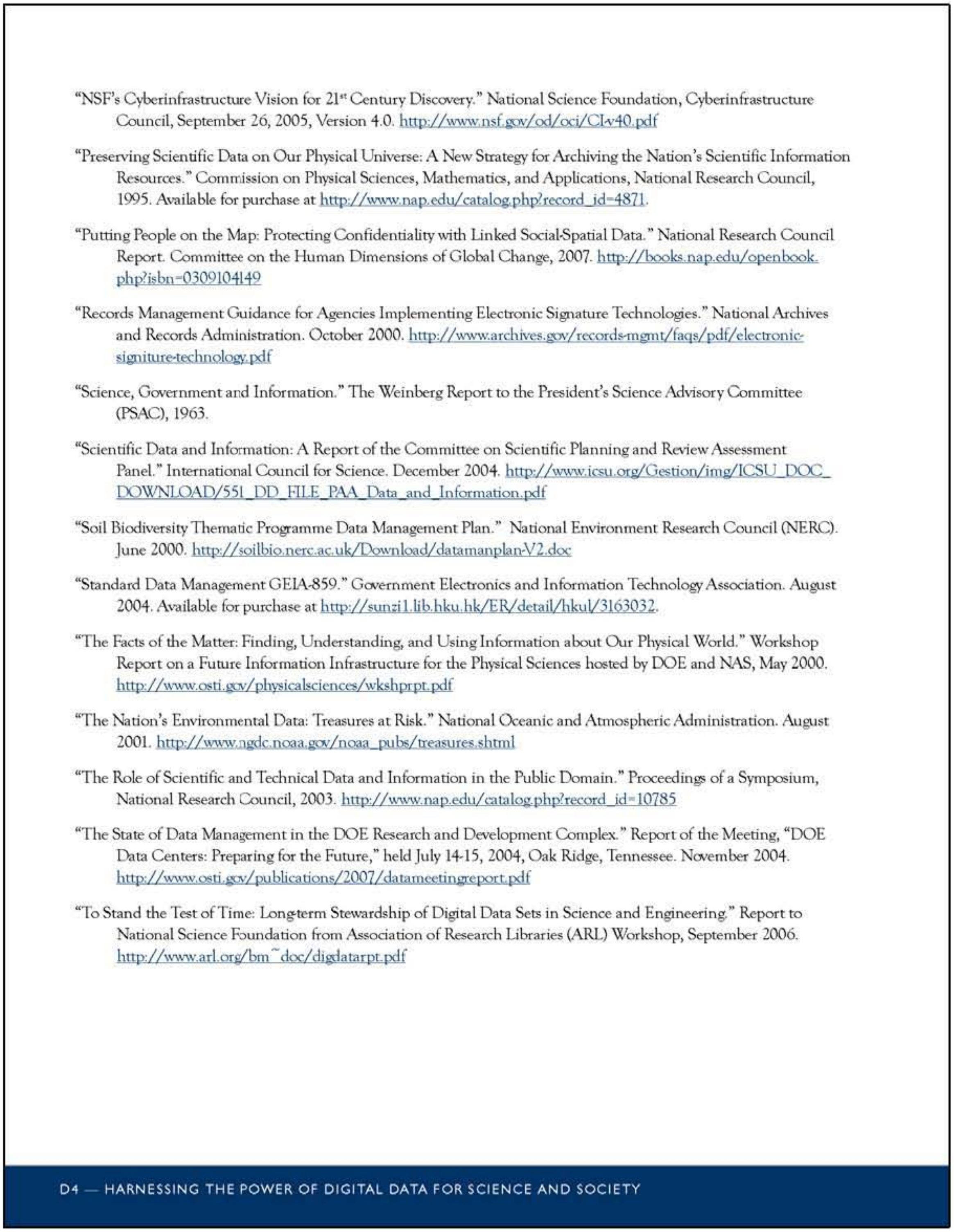
"Environmental Sampling, Analysis and Results, Data Standards Overview of Component Data Standards." Standard No.: EX 000001.0. Environmental Data Standards Council. January 2006. Standard No. EX 000007.1. January 2006.

"Federal Enterprise Architecture Records Management Profile." Version 1.0. Office of Management and Budget. December 2005. http://www.archives.gov/records-mgmt/pdf/rm-profile.pdf.

"Implementation Guide for Data Management GEIA-HB-859. " January 2006. Available for purchase at http://www.techstreet.com/cgi-bin/detail?product_id=1256205.

"Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." National Science Board, September 2005. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

"National Science Foundation Investing in America's Future." Strategic Plan FY 2006-2011, September 2006. http://www.nsf.gov/pubs/2006/nsf0648/NSF-0648.pdf



"NSF's Cyberinfrastructure Vision for 21st Century Discovery." National Science Foundation, Cyberinfrastructure Council, September 26, 2005, Version 4.0. http://www.nsf.gov/od/oci/CI-v40.pdf

"Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources." Commission on Physical Sciences, Mathematics, and Applications, National Research Council, 1995. Available for purchase at http://www.nap.edu/catalog.php?record_id=4871.

"Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data." National Research Council Report. Committee on the Human Dimensions of Global Change, 2007. http://books.nap.edu/openbook.php?isbn=0309104149

"Records Management Guidance for Agencies Implementing Electronic Signature Technologies." National Archives and Records Administration. October 2000. http://www.archives.gov/records-mgmt/faqs/pdf/electronic-signiture-technology.pdf

"Science, Government and Information." The Weinberg Report to the President's Science Advisory Committee (PSAC), 1963.

"Scientific Data and Information: A Report of the Committee on Scientific Planning and Review Assessment Panel." International Council for Science. December 2004. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

"Soil Biodiversity Thematic Programme Data Management Plan." National Environment Research Council (NERC). June 2000. http://soilbio.nerc.ac.uk/Download/datamanplan-V2.doc

"Standard Data Management GEIA-859." Government Electronics and Information Technology Association. August 2004. Available for purchase at http://sunzi1.lib.hku.hk/ER/detail/hkul/3163032.

"The Facts of the Matter: Finding, Understanding, and Using Information about Our Physical World." Workshop Report on a Future Information Infrastructure for the Physical Sciences hosted by DOE and NAS, May 2000. http://www.osti.gov/physicalsciences/wkshprpt.pdf

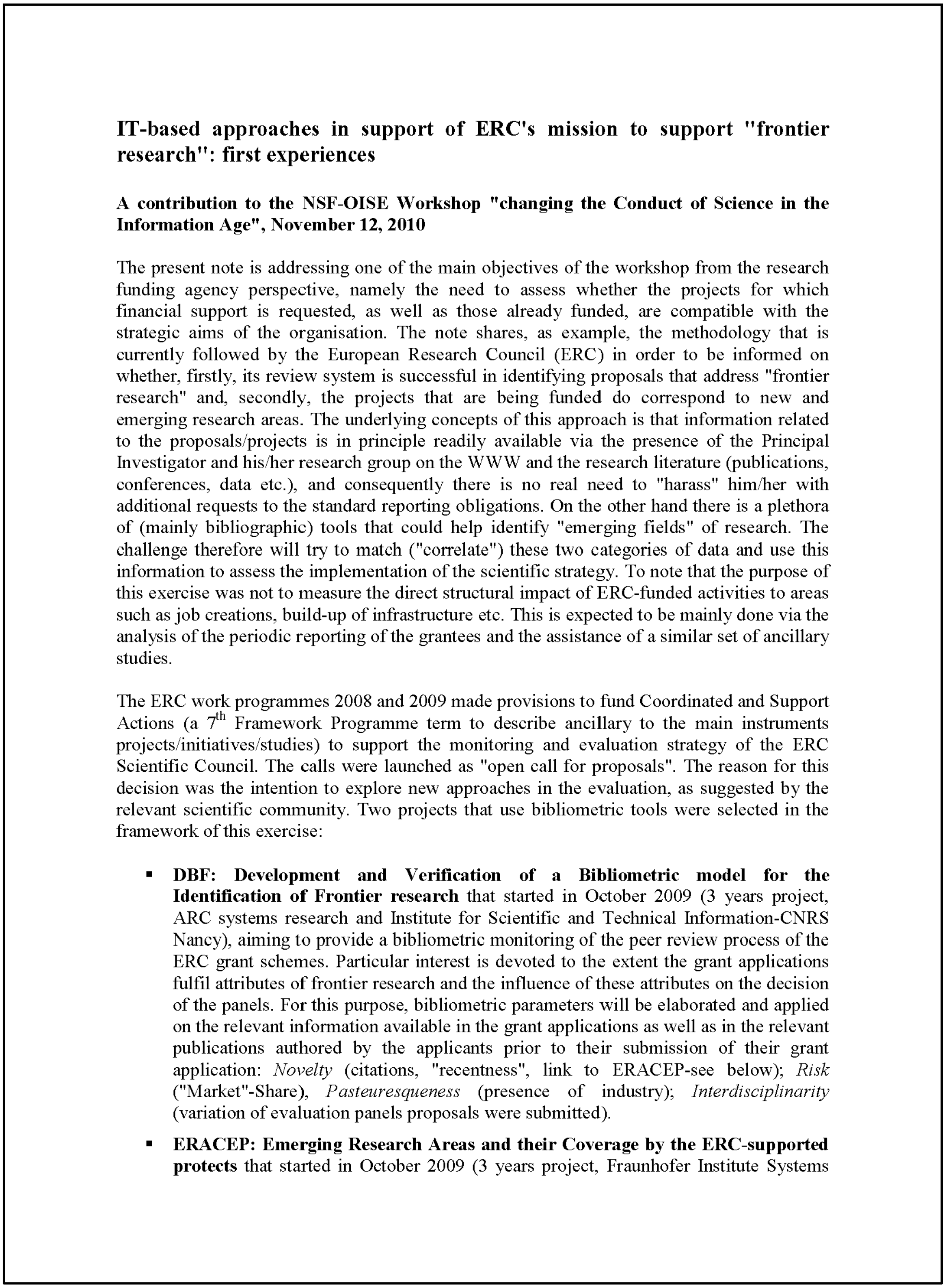
"The Nation's Environmental Data: Treasures at Risk." National Oceanic and Atmospheric Administration. August 2001. http://www.ngdc.noaa.gov/noaa_pubs/treasures.shtml

"The Role of Scientific and Technical Data and Information in the Public Domain." Proceedings of a Symposium, National Research Council, 2003. http://www.nap.edu/catalog.php?record_id=10785

"The State of Data Management in the DOE Research and Development Complex." Report of the Meeting, "DOE Data Centers: Preparing for the Future," held July 14-15, 2004, Oak Ridge, Tennessee. November 2004. http://www.osti.gov/publications/2007/datameetingreport.pdf

"To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering." Report to National Science Foundation from Association of Research Libraries (ARL) Workshop, September 2006. http://www.arl.org/bm~doc/digdatarpt.pdf

Papazoglou, Theodore, "IT-Based Approaches in Support of ERC's Mission to Support 'Frontier Research': First Experiences"



IT-based approaches in support of ERC's mission to support "frontier research": first experiences

A contribution to the NSF-OISE Workshop "changing the Conduct of Science in the Information Age", November 12, 2010

The present note is addressing one of the main objectives of the workshop from the research funding agency perspective, namely the need to assess whether the projects for which financial support is requested, as well as those already funded, are compatible with the strategic aims of the organisation. The note shares, as example, the methodology that is currently followed by the European Research Council (ERC) in order to be informed on whether, firstly, its review system is successful in identifying proposals that address "frontier research" and, secondly, the projects that are being funded do correspond to new and emerging research areas. The underlying concepts of this approach is that information related to the proposals/projects is in principle readily available via the presence of the Principal Investigator and his/her research group on the WWW and the research literature (publications, conferences, data etc.), and consequently there is no real need to "harass" him/her with additional requests to the standard reporting obligations. On the other hand there is a plethora of (mainly bibliographic) tools that could help identify "emerging fields" of research. The challenge therefore will try to match ("correlate") these two categories of data and use this information to assess the implementation of the scientific strategy. To note that the purpose of this exercise was not to measure the direct structural impact of ERC-funded activities to areas such as job creations, build-up of infrastructure etc. This is expected to be mainly done via the analysis of the periodic reporting of the grantees and the assistance of a similar set of ancillary studies.

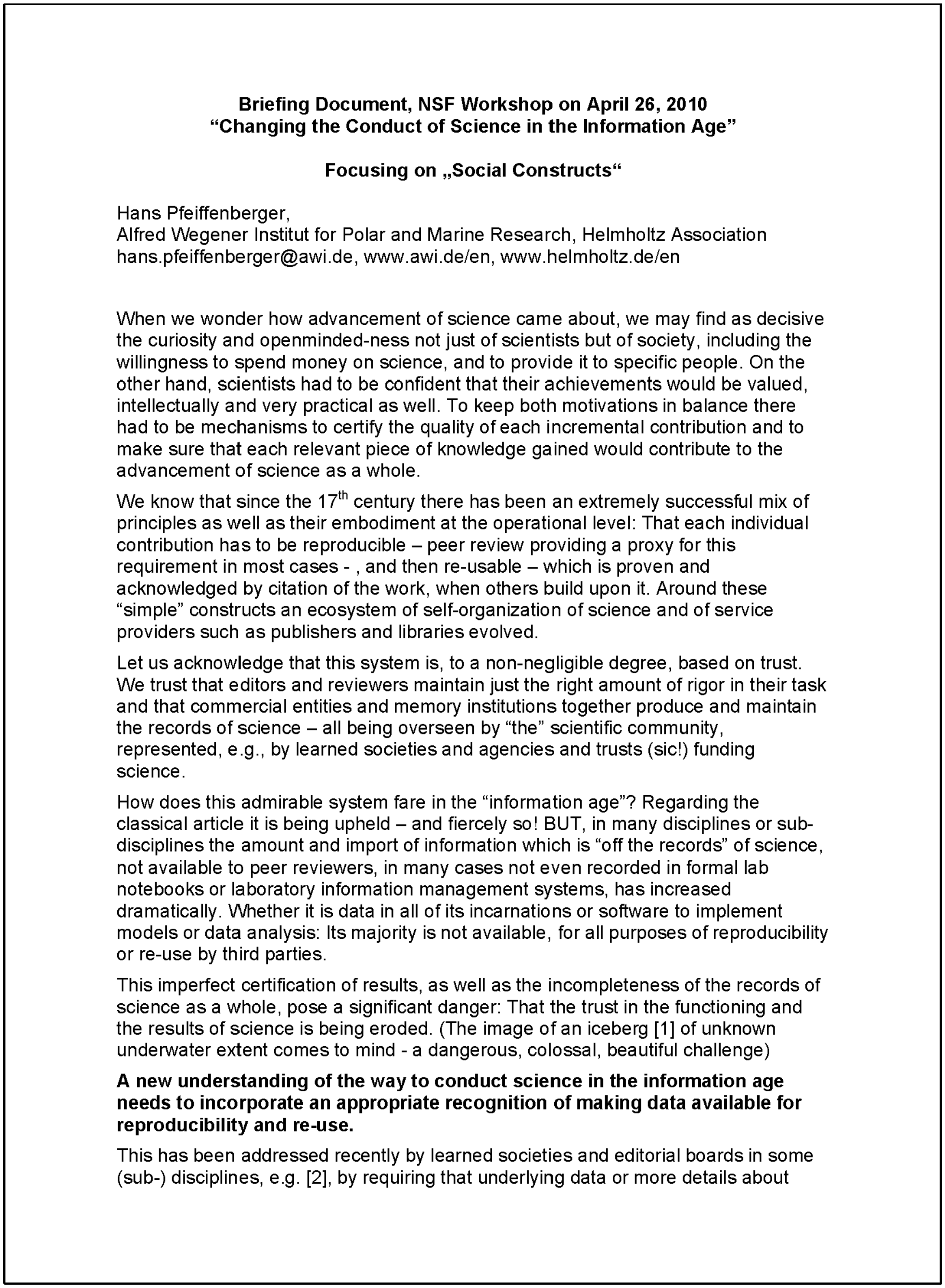
The ERC work programmes 2008 and 2009 made provisions to fund Coordinated and Support Actions (a 7th Framework Programme term to describe ancillary to the main instruments projects/initiatives/studies) to support the monitoring and evaluation strategy of the ERC Scientific Council. The calls were launched as "open call for proposals". The reason for this decision was the intention to explore new approaches in the evaluation, as suggested by the relevant scientific community. Two projects that use bibliometric tools were selected in the framework of this exercise:

- **DBF: Development and Verification of a Bibliometric model for the Identification of Frontier research** that started in October 2009 (3 years project, ARC systems research and Institute for Scientific and Technical Information-CNRS Nancy), aiming to provide a bibliometric monitoring of the peer review process of the ERC grant schemes. Particular interest is devoted to the extent the grant applications fulfil attributes of frontier research and the influence of these attributes on the decision of the panels. For this purpose, bibliometric parameters will be elaborated and applied on the relevant information available in the grant applications as well as in the relevant publications authored by the applicants prior to their submission of their grant application: *Novelty* (citations, "recentness", link to ERACEP-see below); *Risk* ("Market"-Share), *Pasteuresqueness* (presence of industry); *Interdisciplinarity* (variation of evaluation panels proposals were submitted).
- **ERACEP: Emerging Research Areas and their Coverage by the ERC-supported protects** that started in October 2009 (3 years project, Fraunhofer Institute Systems

> and Innovation research and Leuven University – Faculty of business and economics), attempting to identify emerging research areas and analyse to what extent the ERC grants cover and contribute to these research areas. It intends to investigate how ERC is performing in respect to its basic mission: "stimulate scientific excellence by supporting and encouraging the very best, truly creative scientists, scholars and engineers to be adventurous and take risks in their research". The project uses the following methodology (extract from ERACEP's 1st periodic report): "*A set of ISI Subject Categories in the sciences, social sciences and humanities with remarkable growth in the last decade are defined. Twenty categories have been selected to undergo further structural analysis; the objective is to identify new and/or emerging topics within these subject matters. In particular, 13 fields have been selected from the sciences, 5 from the social sciences and 3 from the humanities. The underlying data have been retrieved from Thomson Reuters' Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI) for the period 1998-2007. Cluster analysis in bibliometrics is traditionally based on both citation links (bibliographic coupling, cross-citation, co-citation analysis) and textual links (co-word analysis, term representation). Both approaches have advantages and shortcomings. The main advantage of citation-based methods is their discriminative power. This is contrasted by a serious disadvantage: Citation-link matrices are extremely sparse and citation-cased methods tend to "underestimate" links among documents. Furthermore, citation links generate binary measures which are based on value 0 or 1 according as there is a citation link between two documents or not. By contrast, text-based measures are based on term frequencies in documents, which as such provide a natural weight underlying the similarity/distance measures used for the analysis. Link matrices are furthermore less sparse than their counterparts in the citation space. These advantages are cancelled out by two serious problems: The lower discriminative power, which results in "overestimating" links among documents and the dimensionality problem. At least the latter problem can be compensated by Singular Value Decomposition (SVD) or directly by Latent Semantic Indexing (LSI), which uses the first-mentioned algorithm*".

Some initial technical obstacles as well as legal constraints will be reported during the workshop in order to assist the participants to appreciate that even when "enabling technologies" are in the disposition of a research funding agency, the challenges are still significant.

Pfeiffenberger, Hans, “Focusing on Social Constructs”



Briefing Document, NSF Workshop on April 26, 2010
“Changing the Conduct of Science in the Information Age”

Focusing on „Social Constructs“

Hans Pfeiffenberger,
Alfred Wegener Institut for Polar and Marine Research, Helmholtz Association
hans.pfeiffenberger@awi.de, www.awi.de/en, www.helmholtz.de/en

When we wonder how advancement of science came about, we may find as decisive the curiosity and openminded-ness not just of scientists but of society, including the willingness to spend money on science, and to provide it to specific people. On the other hand, scientists had to be confident that their achievements would be valued, intellectually and very practical as well. To keep both motivations in balance there had to be mechanisms to certify the quality of each incremental contribution and to make sure that each relevant piece of knowledge gained would contribute to the advancement of science as a whole.

We know that since the 17th century there has been an extremely successful mix of principles as well as their embodiment at the operational level: That each individual contribution has to be reproducible – peer review providing a proxy for this requirement in most cases - , and then re-usable – which is proven and acknowledged by citation of the work, when others build upon it. Around these “simple” constructs an ecosystem of self-organization of science and of service providers such as publishers and libraries evolved.

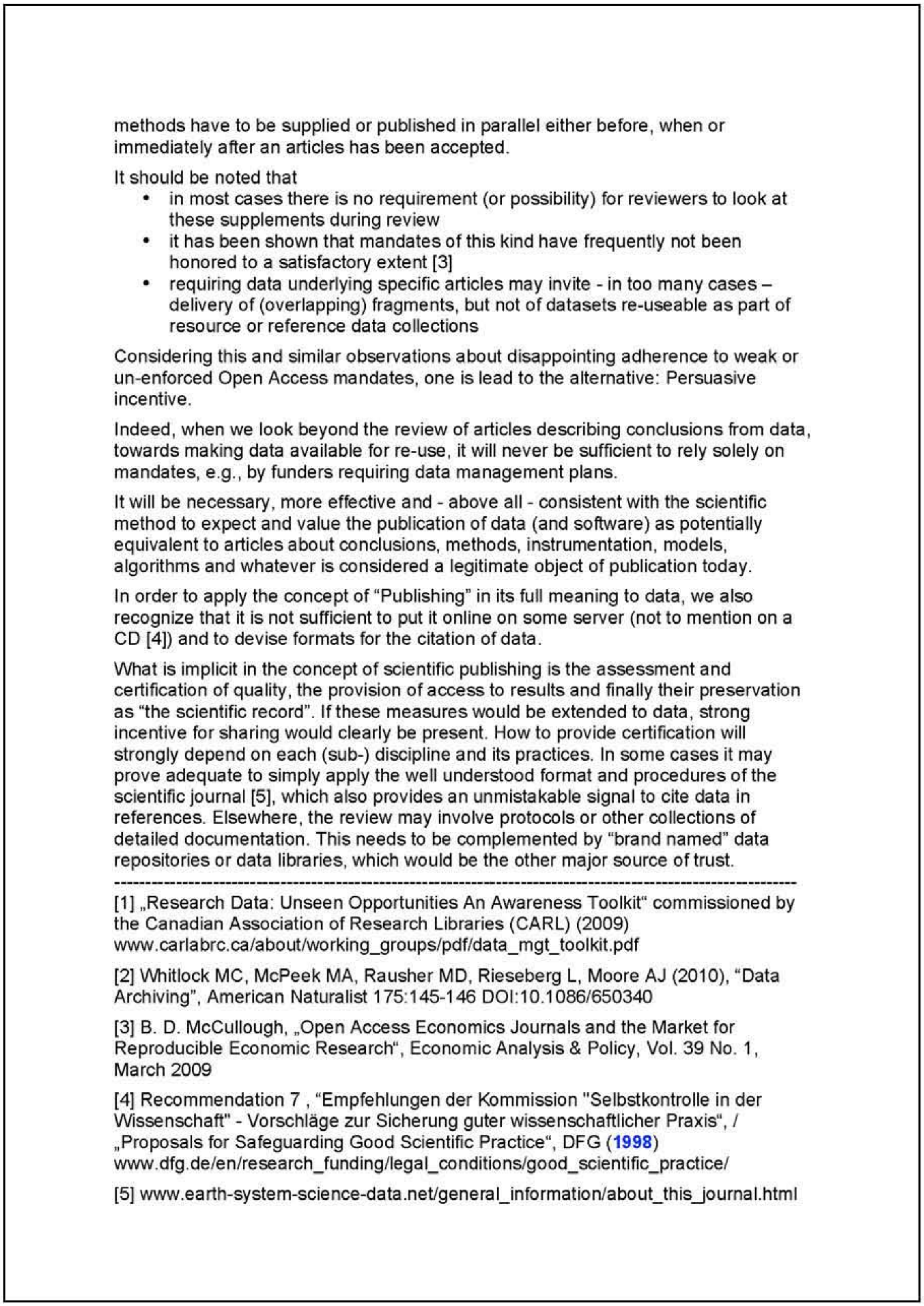
Let us acknowledge that this system is, to a non-negligible degree, based on trust. We trust that editors and reviewers maintain just the right amount of rigor in their task and that commercial entities and memory institutions together produce and maintain the records of science – all being overseen by “the” scientific community, represented, e.g., by learned societies and agencies and trusts (sic!) funding science.

How does this admirable system fare in the “information age”? Regarding the classical article it is being upheld – and fiercely so! BUT, in many disciplines or sub-disciplines the amount and import of information which is “off the records” of science, not available to peer reviewers, in many cases not even recorded in formal lab notebooks or laboratory information management systems, has increased dramatically. Whether it is data in all of its incarnations or software to implement models or data analysis: Its majority is not available, for all purposes of reproducibility or re-use by third parties.

This imperfect certification of results, as well as the incompleteness of the records of science as a whole, pose a significant danger: That the trust in the functioning and the results of science is being eroded. (The image of an iceberg [1] of unknown underwater extent comes to mind - a dangerous, colossal, beautiful challenge)

A new understanding of the way to conduct science in the information age needs to incorporate an appropriate recognition of making data available for reproducibility and re-use.

This has been addressed recently by learned societies and editorial boards in some (sub-) disciplines, e.g. [2], by requiring that underlying data or more details about



methods have to be supplied or published in parallel either before, when or immediately after an articles has been accepted.

It should be noted that

- in most cases there is no requirement (or possibility) for reviewers to look at these supplements during review
- it has been shown that mandates of this kind have frequently not been honored to a satisfactory extent [3]
- requiring data underlying specific articles may invite - in too many cases – delivery of (overlapping) fragments, but not of datasets re-useable as part of resource or reference data collections

Considering this and similar observations about disappointing adherence to weak or un-enforced Open Access mandates, one is lead to the alternative: Persuasive incentive.

Indeed, when we look beyond the review of articles describing conclusions from data, towards making data available for re-use, it will never be sufficient to rely solely on mandates, e.g., by funders requiring data management plans.

It will be necessary, more effective and - above all - consistent with the scientific method to expect and value the publication of data (and software) as potentially equivalent to articles about conclusions, methods, instrumentation, models, algorithms and whatever is considered a legitimate object of publication today.

In order to apply the concept of "Publishing" in its full meaning to data, we also recognize that it is not sufficient to put it online on some server (not to mention on a CD [4]) and to devise formats for the citation of data.

What is implicit in the concept of scientific publishing is the assessment and certification of quality, the provision of access to results and finally their preservation as "the scientific record". If these measures would be extended to data, strong incentive for sharing would clearly be present. How to provide certification will strongly depend on each (sub-) discipline and its practices. In some cases it may prove adequate to simply apply the well understood format and procedures of the scientific journal [5], which also provides an unmistakable signal to cite data in references. Elsewhere, the review may involve protocols or other collections of detailed documentation. This needs to be complemented by "brand named" data repositories or data libraries, which would be the other major source of trust.

---

[1] „Research Data: Unseen Opportunities An Awareness Toolkit" commissioned by the Canadian Association of Research Libraries (CARL) (2009) www.carlabrc.ca/about/working_groups/pdf/data_mgt_toolkit.pdf

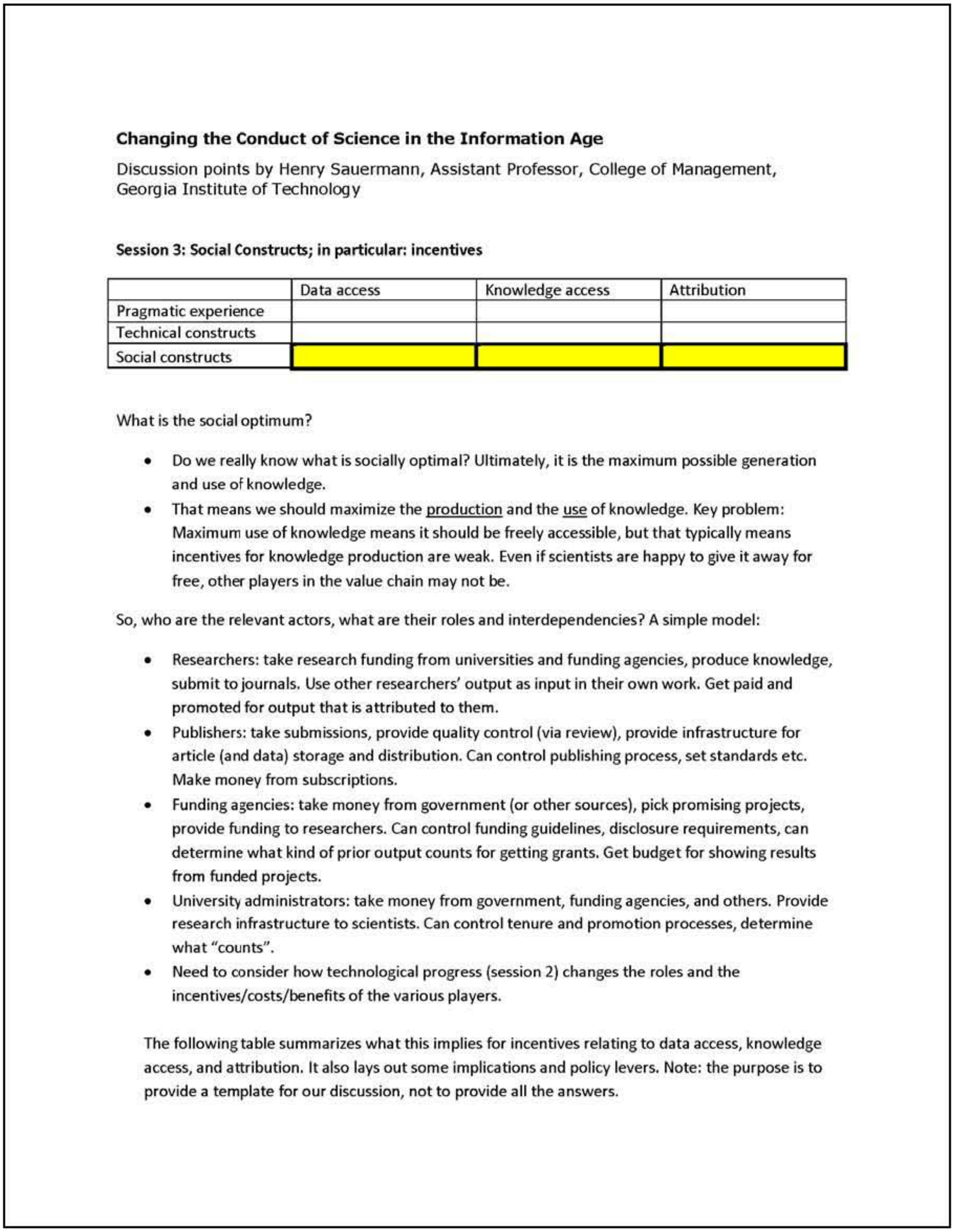
[2] Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010), "Data Archiving", American Naturalist 175:145-146 DOI:10.1086/650340

[3] B. D. McCullough, „Open Access Economics Journals and the Market for Reproducible Economic Research", Economic Analysis & Policy, Vol. 39 No. 1, March 2009

[4] Recommendation 7 , "Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft" - Vorschläge zur Sicherung guter wissenschaftlicher Praxis", / „Proposals for Safeguarding Good Scientific Practice", DFG (**1998**) www.dfg.de/en/research_funding/legal_conditions/good_scientific_practice/

[5] www.earth-system-science-data.net/general_information/about_this_journal.html

Sauermann, Henry, “Discussion Points for Session 3: Social Constructs; in Particular: Incentives”



Changing the Conduct of Science in the Information Age

Discussion points by Henry Sauermann, Assistant Professor, College of Management, Georgia Institute of Technology

Session 3: Social Constructs; in particular: incentives

	Data access	Knowledge access	Attribution
Pragmatic experience			
Technical constructs			
Social constructs	🟨	🟨	🟨

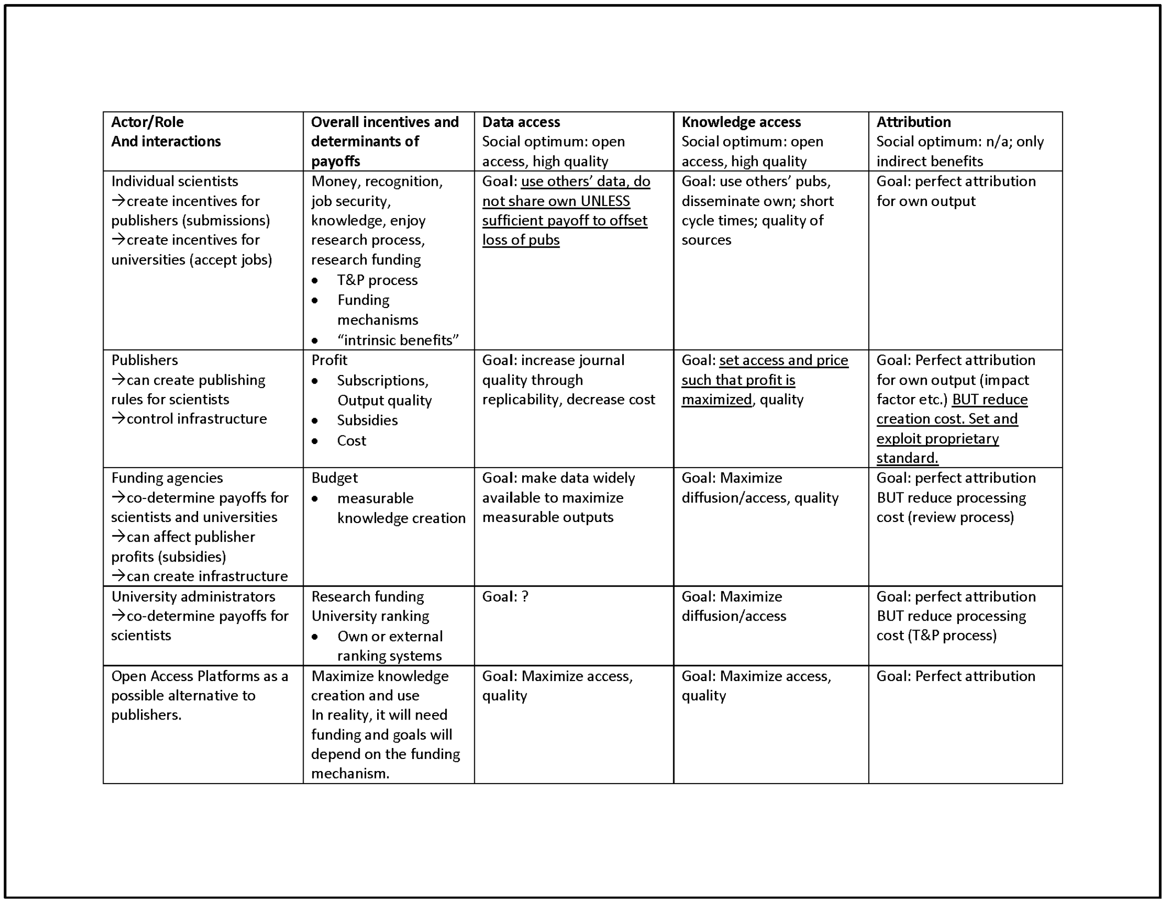
What is the social optimum?

- Do we really know what is socially optimal? Ultimately, it is the maximum possible generation and use of knowledge.
- That means we should maximize the production and the use of knowledge. Key problem: Maximum use of knowledge means it should be freely accessible, but that typically means incentives for knowledge production are weak. Even if scientists are happy to give it away for free, other players in the value chain may not be.

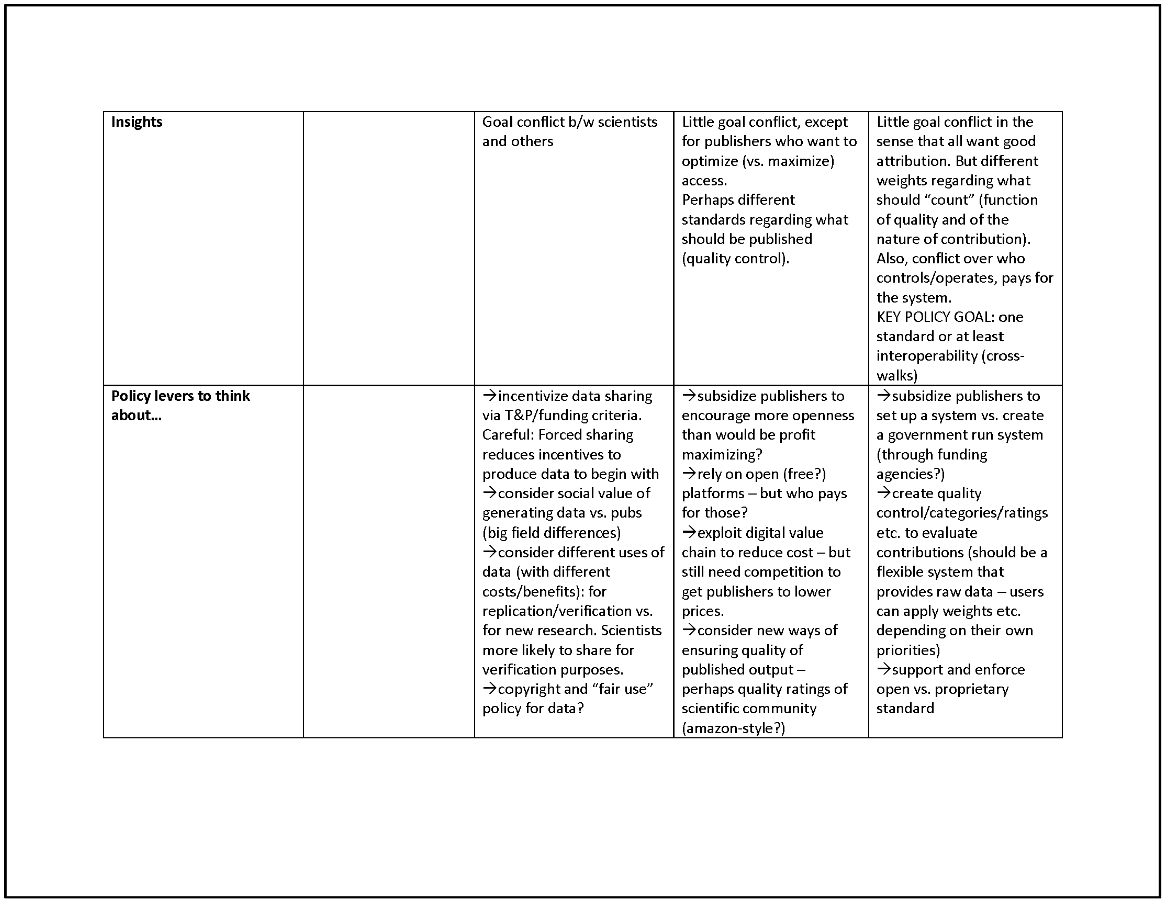
So, who are the relevant actors, what are their roles and interdependencies? A simple model:

- Researchers: take research funding from universities and funding agencies, produce knowledge, submit to journals. Use other researchers' output as input in their own work. Get paid and promoted for output that is attributed to them.
- Publishers: take submissions, provide quality control (via review), provide infrastructure for article (and data) storage and distribution. Can control publishing process, set standards etc. Make money from subscriptions.
- Funding agencies: take money from government (or other sources), pick promising projects, provide funding to researchers. Can control funding guidelines, disclosure requirements, can determine what kind of prior output counts for getting grants. Get budget for showing results from funded projects.
- University administrators: take money from government, funding agencies, and others. Provide research infrastructure to scientists. Can control tenure and promotion processes, determine what “counts”.
- Need to consider how technological progress (session 2) changes the roles and the incentives/costs/benefits of the various players.

The following table summarizes what this implies for incentives relating to data access, knowledge access, and attribution. It also lays out some implications and policy levers. Note: the purpose is to provide a template for our discussion, not to provide all the answers.



Actor/Role **And interactions**	**Overall incentives and determinants of payoffs**	**Data access** Social optimum: open access, high quality	**Knowledge access** Social optimum: open access, high quality	**Attribution** Social optimum: n/a; only indirect benefits
Individual scientists →create incentives for publishers (submissions) →create incentives for universities (accept jobs)	Money, recognition, job security, knowledge, enjoy research process, research funding • T&P process • Funding mechanisms • "intrinsic benefits"	Goal: use others' data, do not share own UNLESS sufficient payoff to offset loss of pubs	Goal: use others' pubs, disseminate own; short cycle times; quality of sources	Goal: perfect attribution for own output
Publishers →can create publishing rules for scientists →control infrastructure	Profit • Subscriptions, Output quality • Subsidies • Cost	Goal: increase journal quality through replicability, decrease cost	Goal: set access and price such that profit is maximized, quality	Goal: Perfect attribution for own output (impact factor etc.) BUT reduce creation cost. Set and exploit proprietary standard.
Funding agencies →co-determine payoffs for scientists and universities →can affect publisher profits (subsidies) →can create infrastructure	Budget • measurable knowledge creation	Goal: make data widely available to maximize measurable outputs	Goal: Maximize diffusion/access, quality	Goal: perfect attribution BUT reduce processing cost (review process)
University administrators →co-determine payoffs for scientists	Research funding University ranking • Own or external ranking systems	Goal: ?	Goal: Maximize diffusion/access	Goal: perfect attribution BUT reduce processing cost (T&P process)
Open Access Platforms as a possible alternative to publishers.	Maximize knowledge creation and use In reality, it will need funding and goals will depend on the funding mechanism.	Goal: Maximize access, quality	Goal: Maximize access, quality	Goal: Perfect attribution



Insights		Goal conflict b/w scientists and others	Little goal conflict, except for publishers who want to optimize (vs. maximize) access. Perhaps different standards regarding what should be published (quality control).	Little goal conflict in the sense that all want good attribution. But different weights regarding what should "count" (function of quality and of the nature of contribution). Also, conflict over who controls/operates, pays for the system. KEY POLICY GOAL: one standard or at least interoperability (cross-walks)
Policy levers to think about…		→incentivize data sharing via T&P/funding criteria. Careful: Forced sharing reduces incentives to produce data to begin with →consider social value of generating data vs. pubs (big field differences) →consider different uses of data (with different costs/benefits): for replication/verification vs. for new research. Scientists more likely to share for verification purposes. →copyright and "fair use" policy for data?	→subsidize publishers to encourage more openness than would be profit maximizing? →rely on open (free?) platforms – but who pays for those? →exploit digital value chain to reduce cost – but still need competition to get publishers to lower prices. →consider new ways of ensuring quality of published output – perhaps quality ratings of scientific community (amazon-style?)	→subsidize publishers to set up a system vs. create a government run system (through funding agencies?) →create quality control/categories/ratings etc. to evaluate contributions (should be a flexible system that provides raw data – users can apply weights etc. depending on their own priorities) →support and enforce open vs. proprietary standard

Schutz, Bernard, "Data Access: Digital Technology and Scientific Communities"

DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1

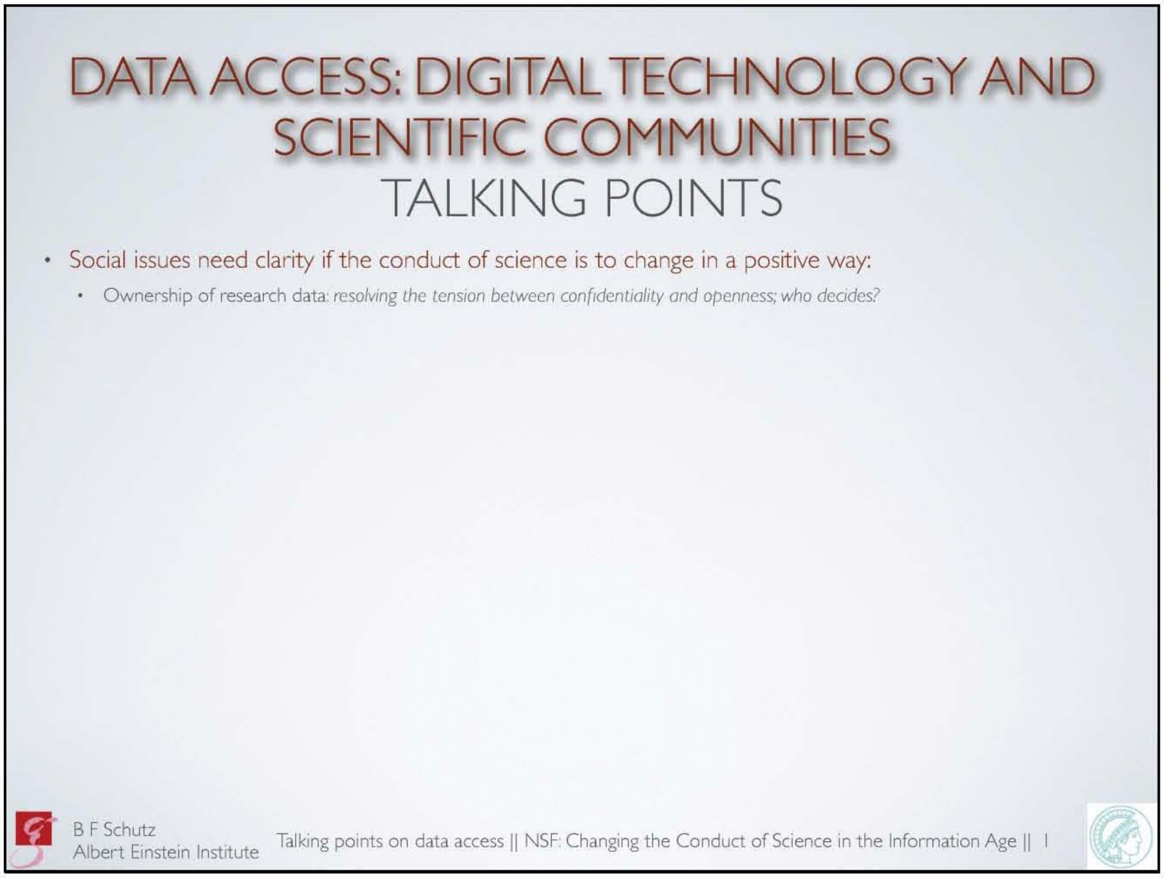
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1

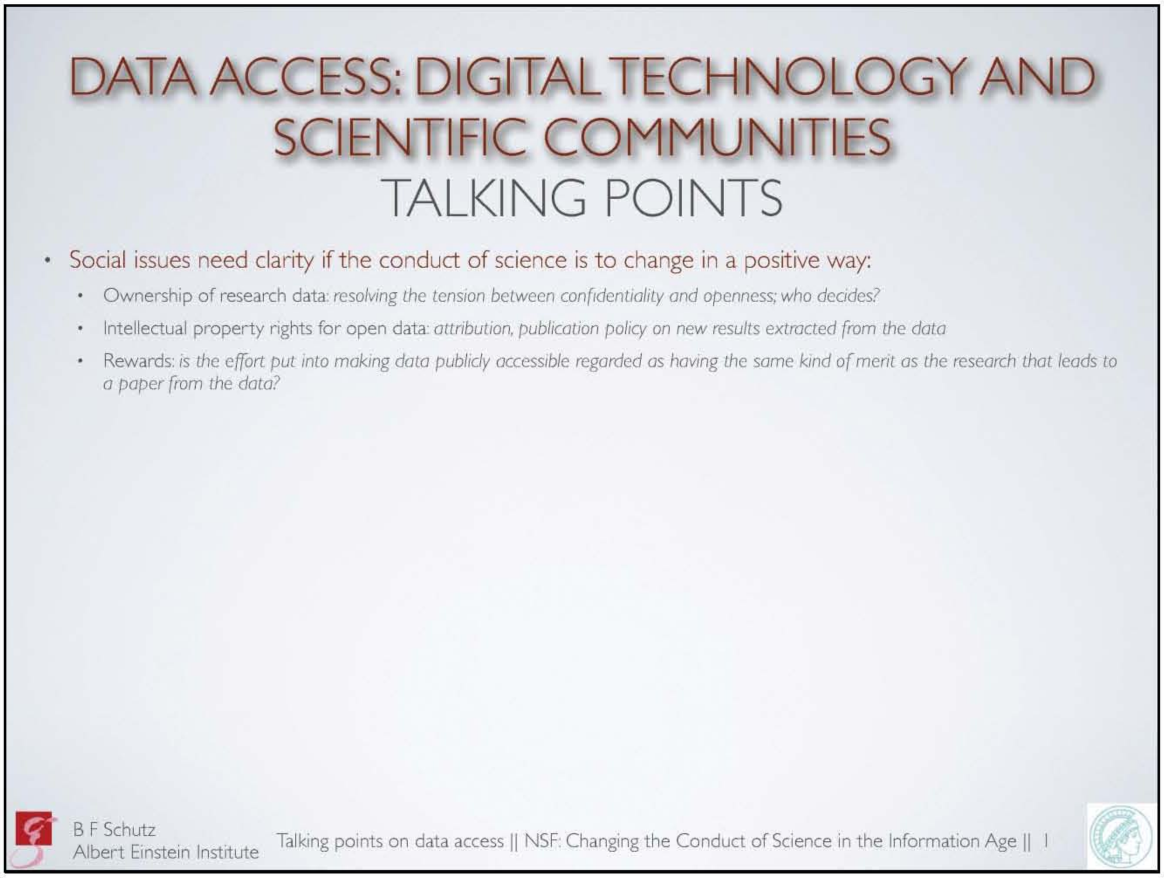
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



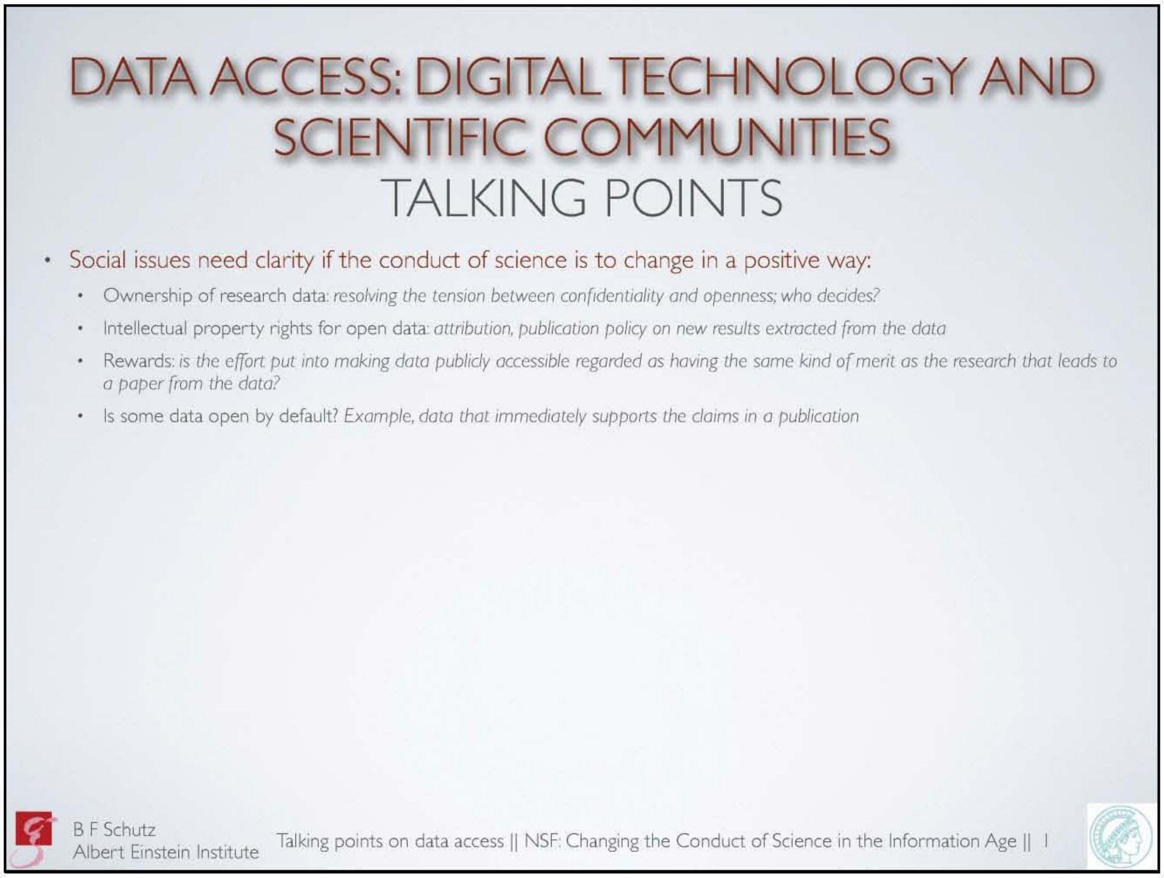
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



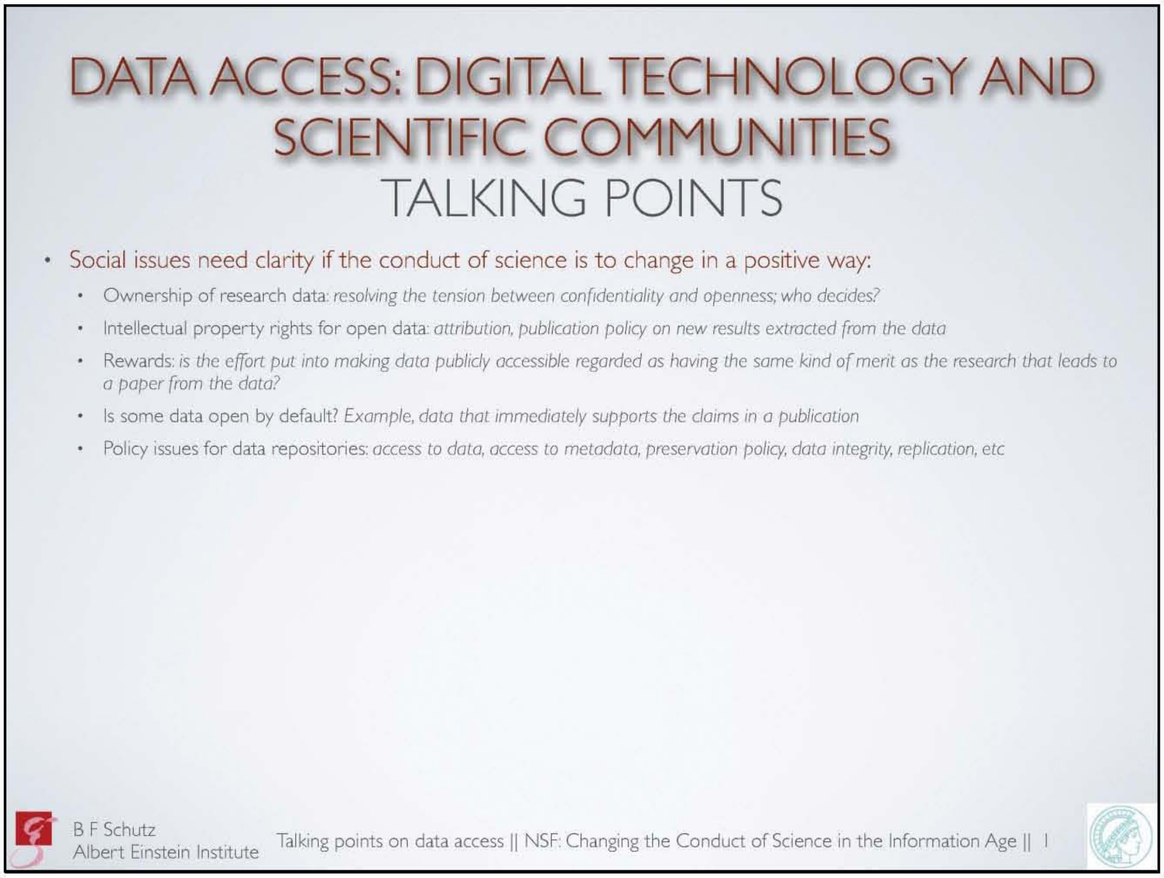
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



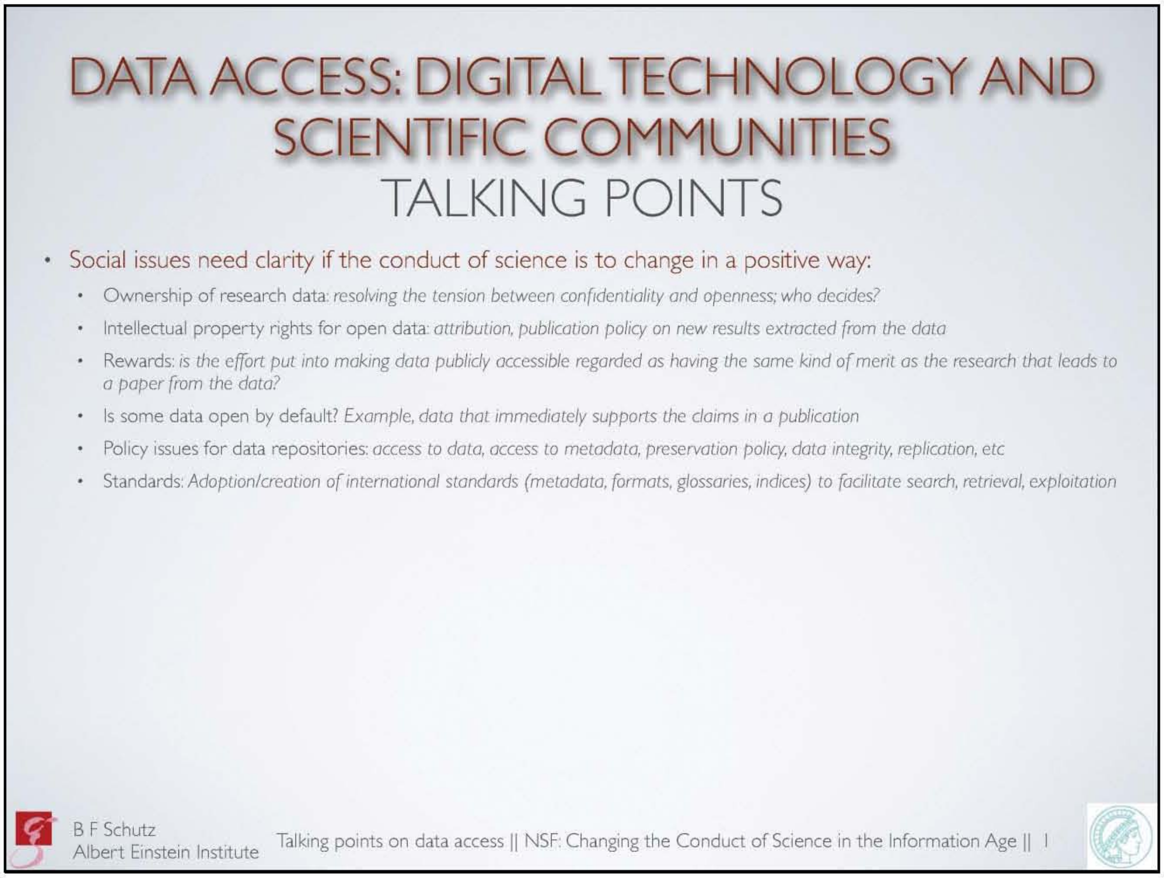
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



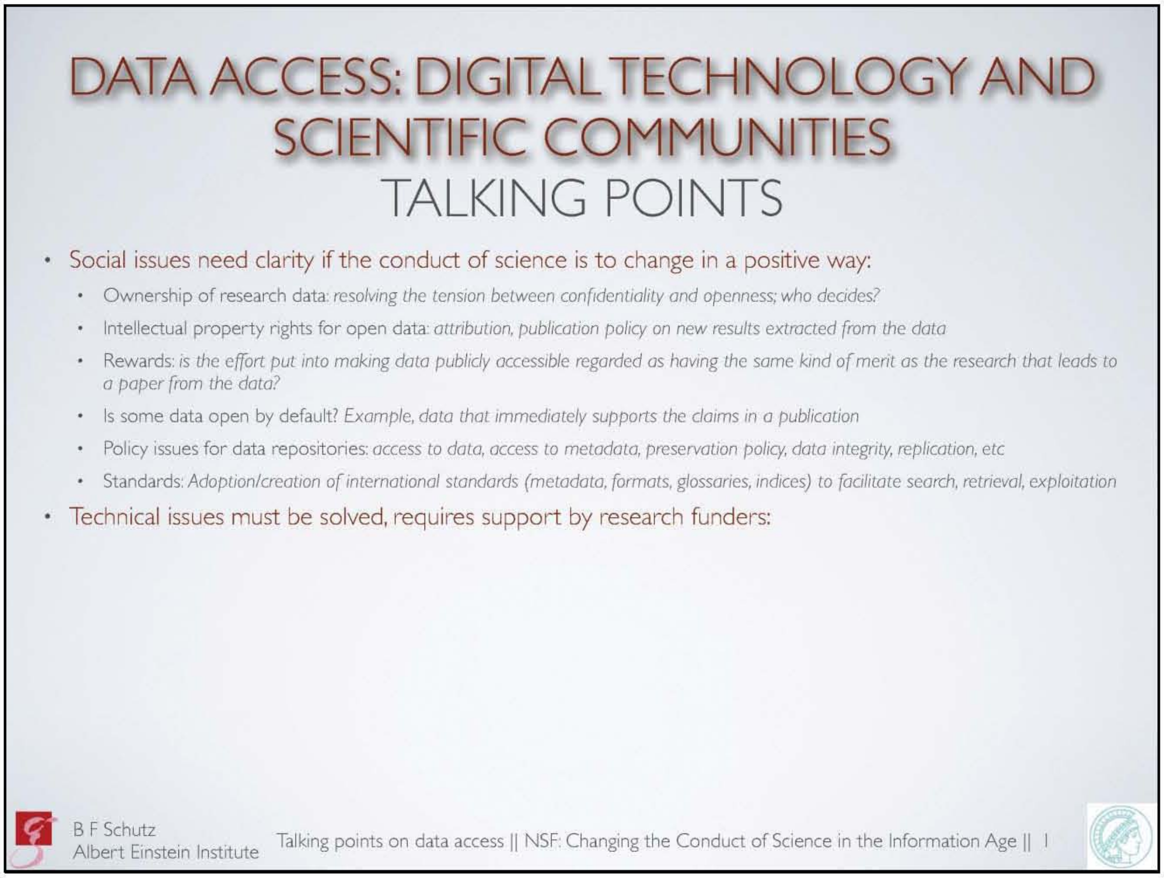
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



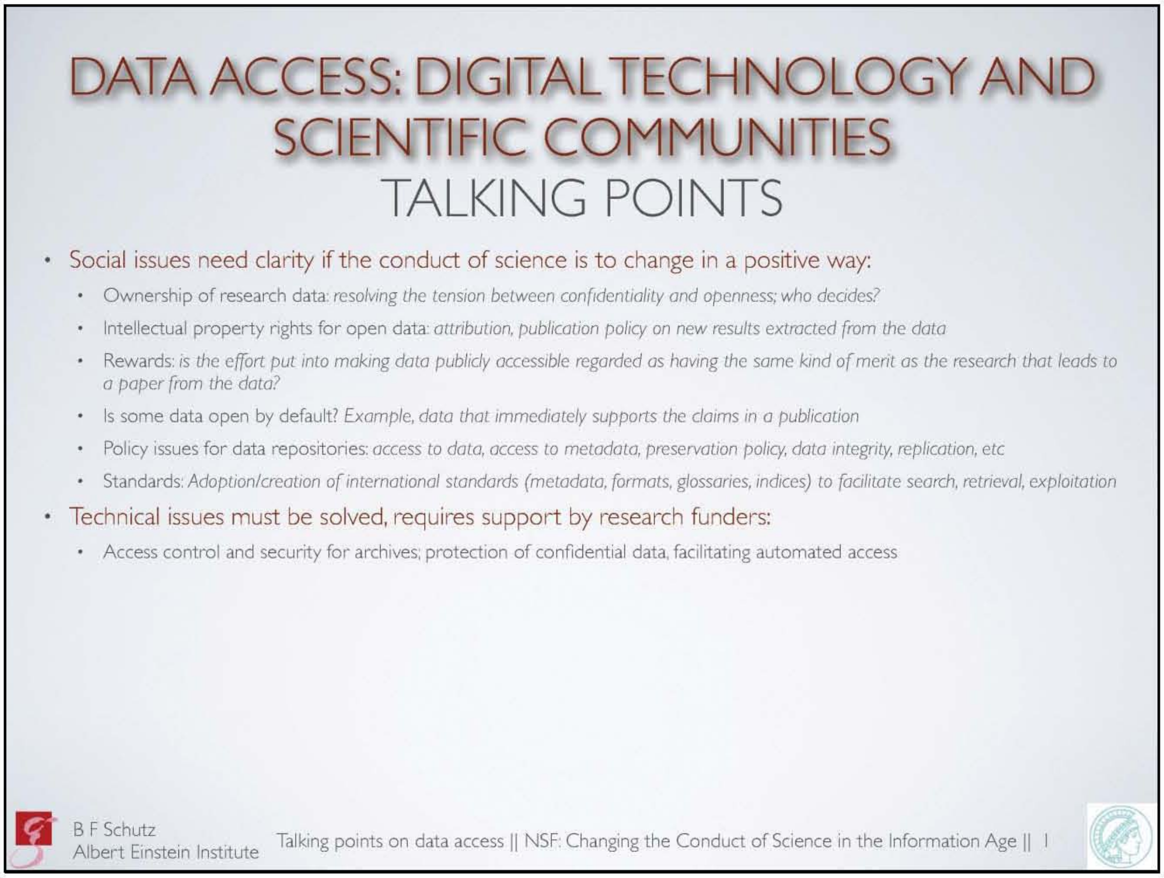
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



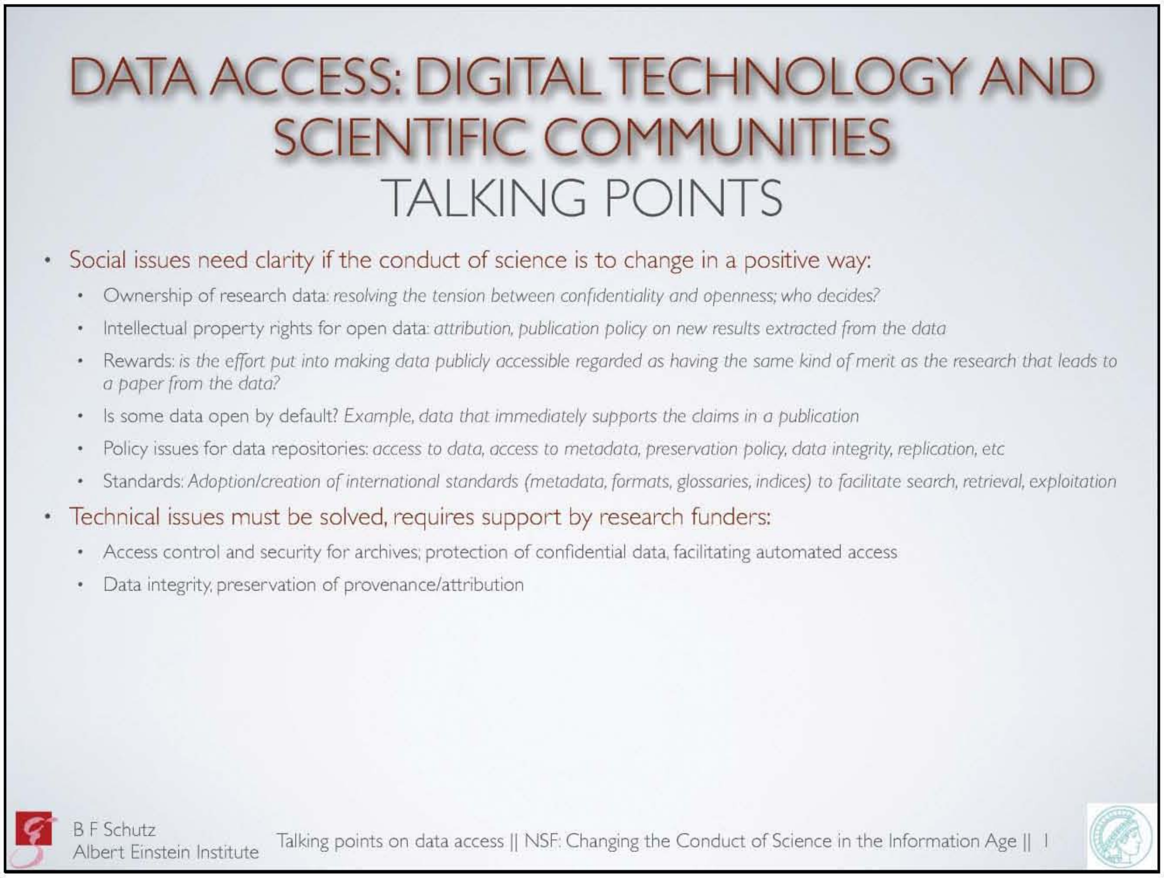
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



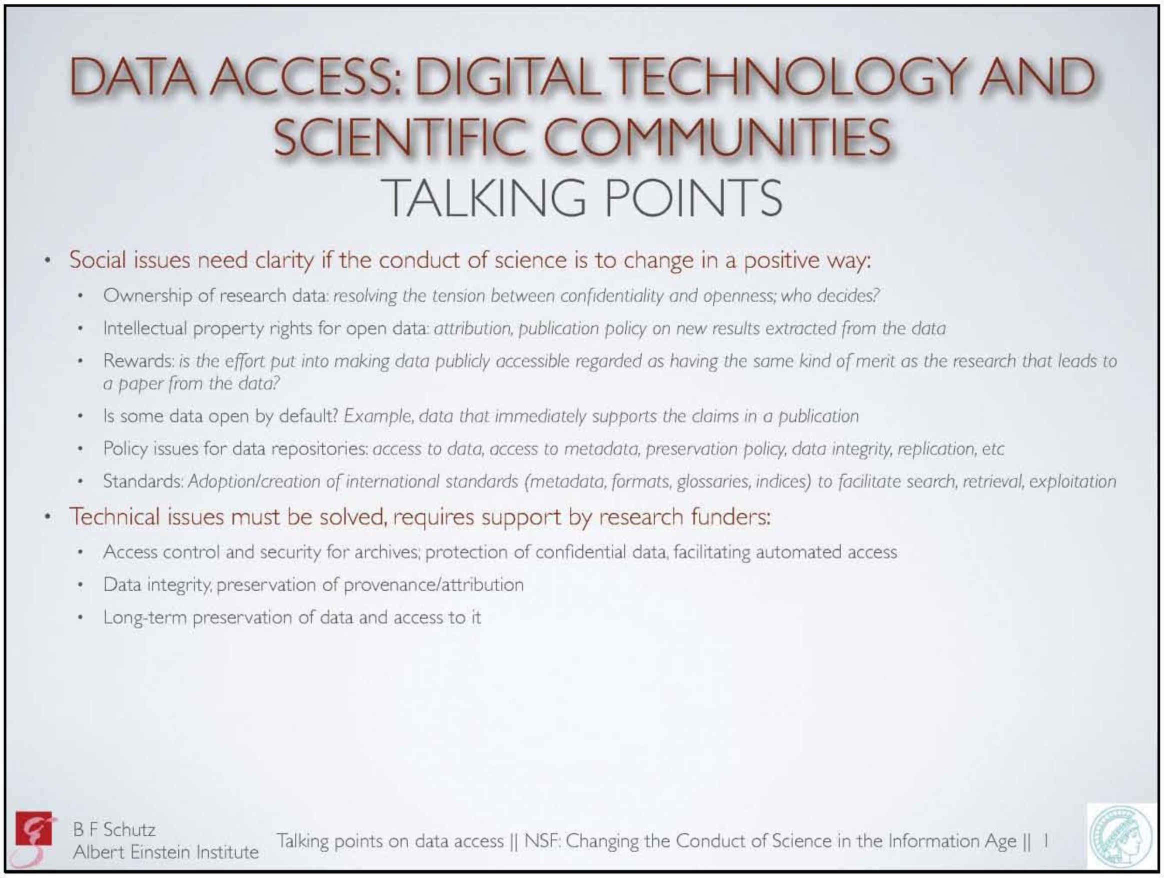
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



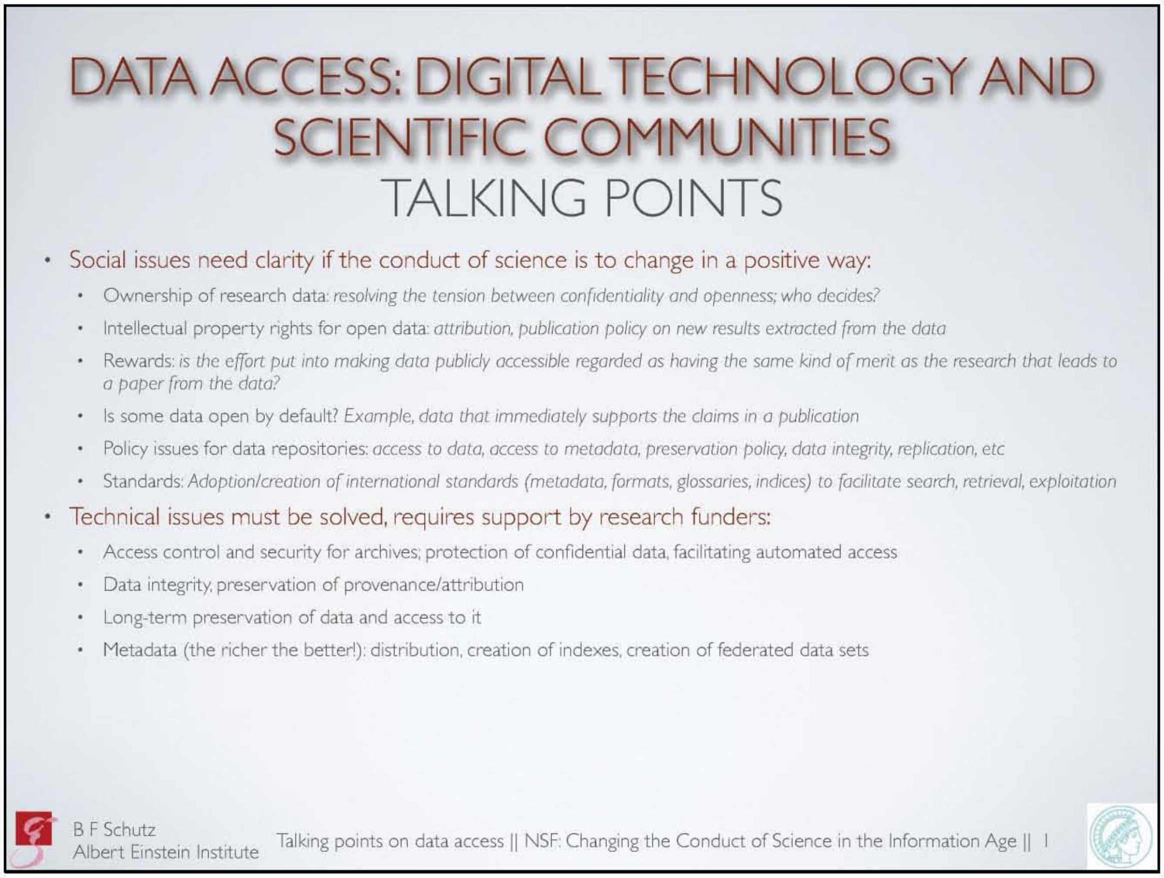
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



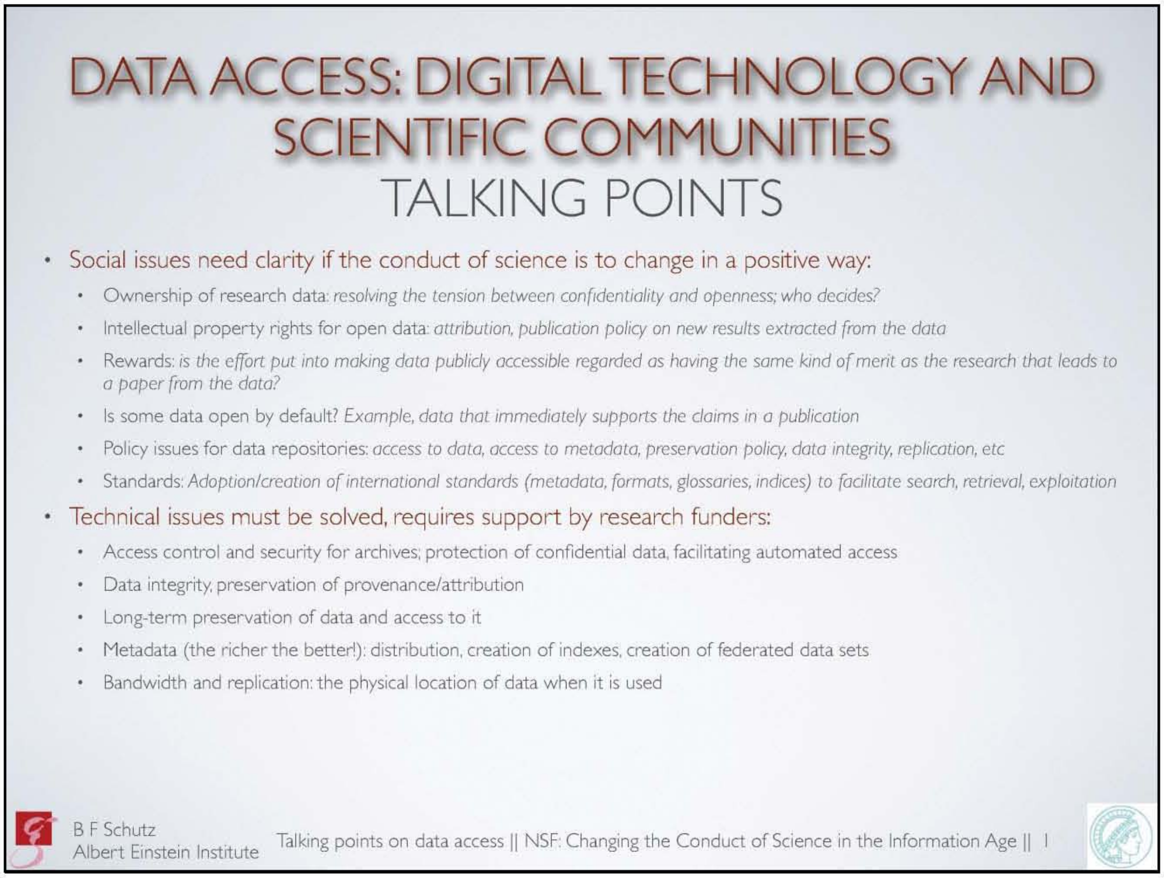
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it
  - Metadata (the richer the better!): distribution, creation of indexes, creation of federated data sets

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



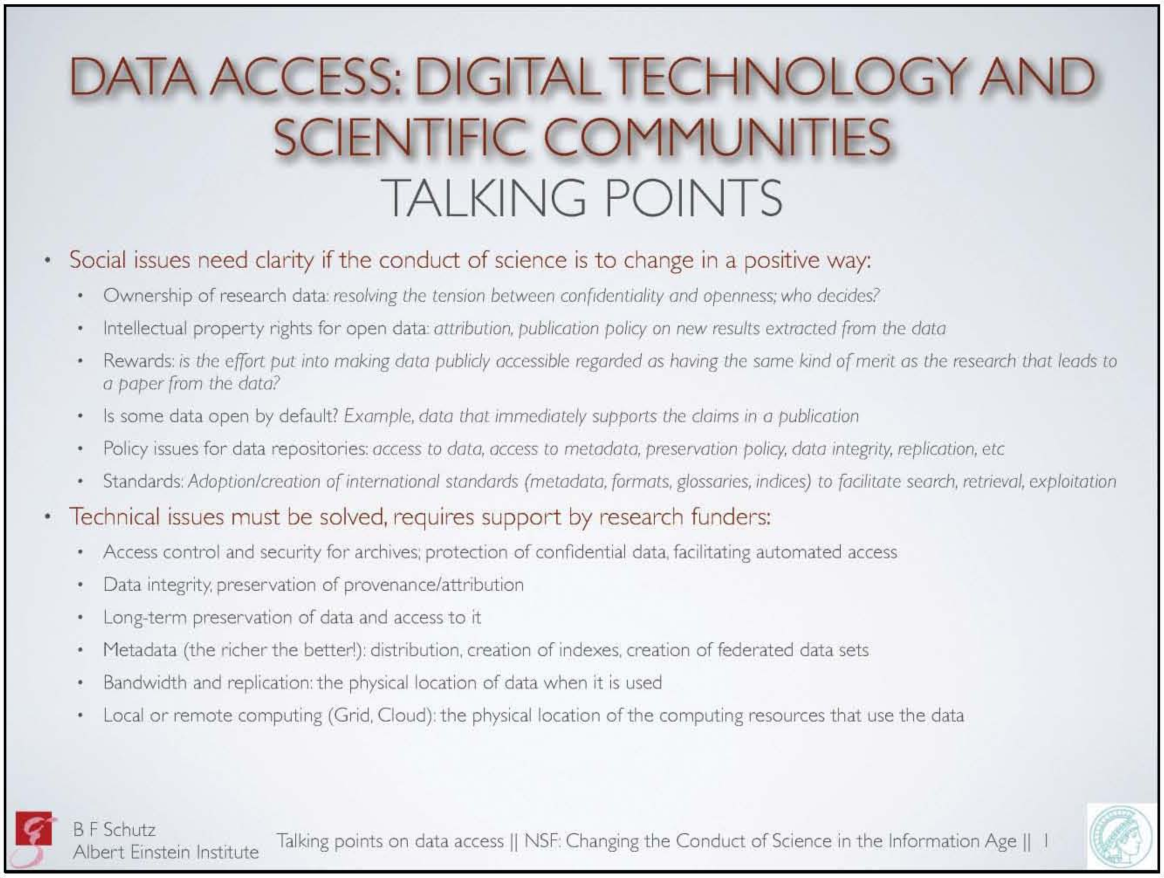
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it
  - Metadata (the richer the better!): distribution, creation of indexes, creation of federated data sets
  - Bandwidth and replication: the physical location of data when it is used

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



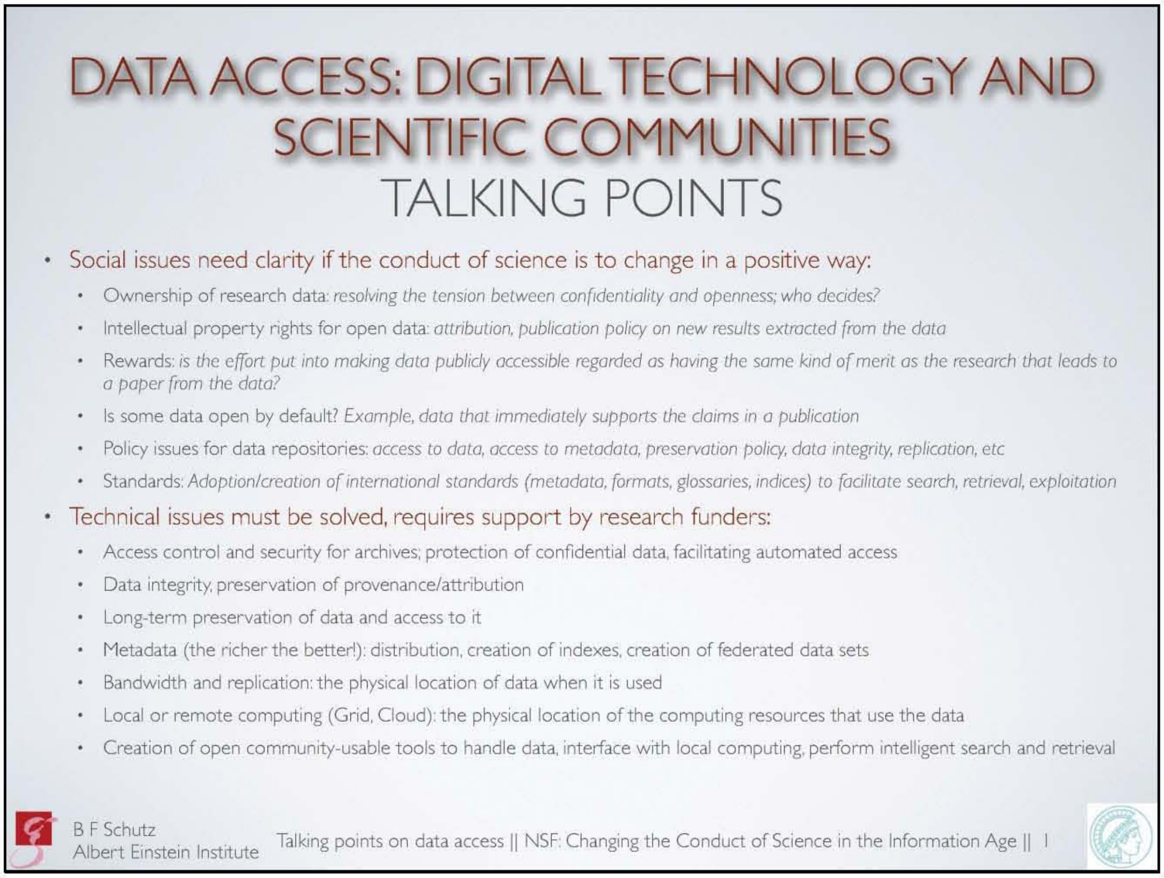
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it
  - Metadata (the richer the better!): distribution, creation of indexes, creation of federated data sets
  - Bandwidth and replication: the physical location of data when it is used
  - Local or remote computing (Grid, Cloud): the physical location of the computing resources that use the data

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



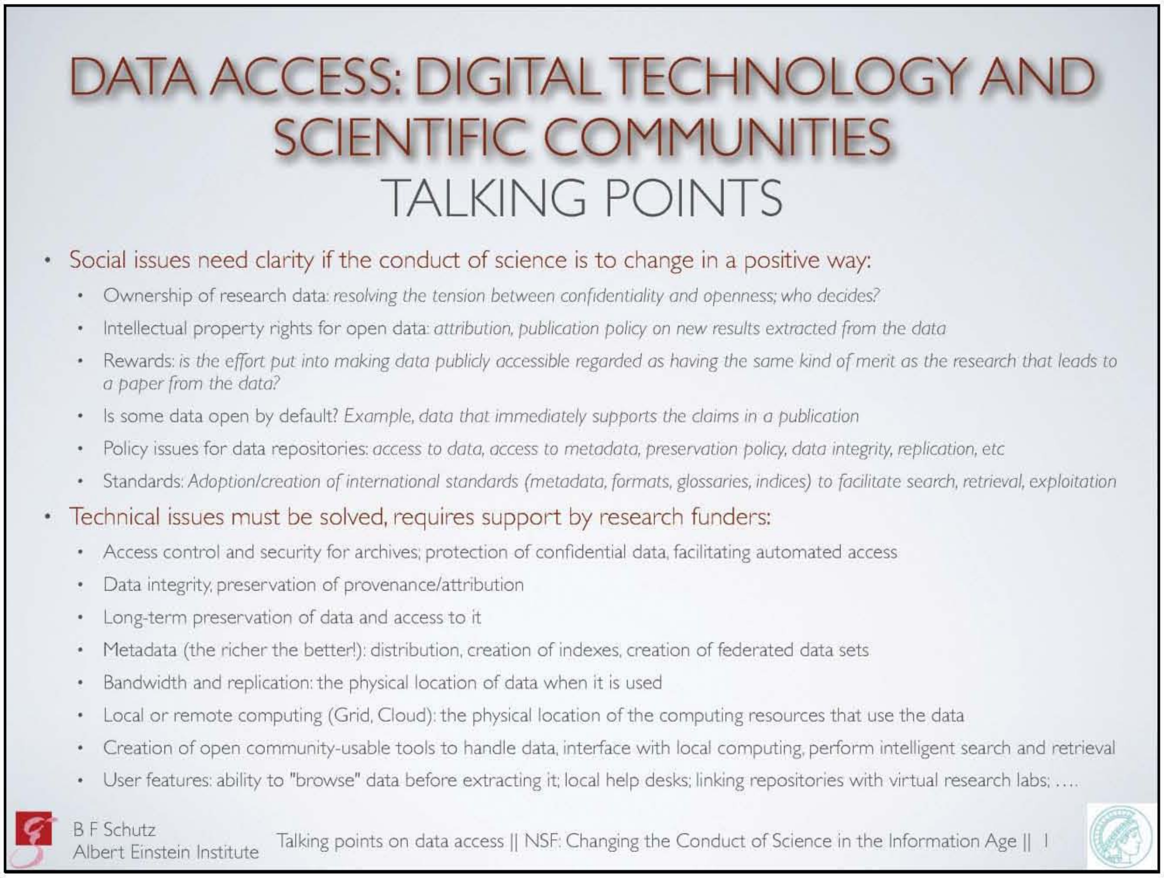
DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it
  - Metadata (the richer the better!): distribution, creation of indexes, creation of federated data sets
  - Bandwidth and replication: the physical location of data when it is used
  - Local or remote computing (Grid, Cloud): the physical location of the computing resources that use the data
  - Creation of open community-usable tools to handle data, interface with local computing, perform intelligent search and retrieval

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1



DATA ACCESS: DIGITAL TECHNOLOGY AND SCIENTIFIC COMMUNITIES

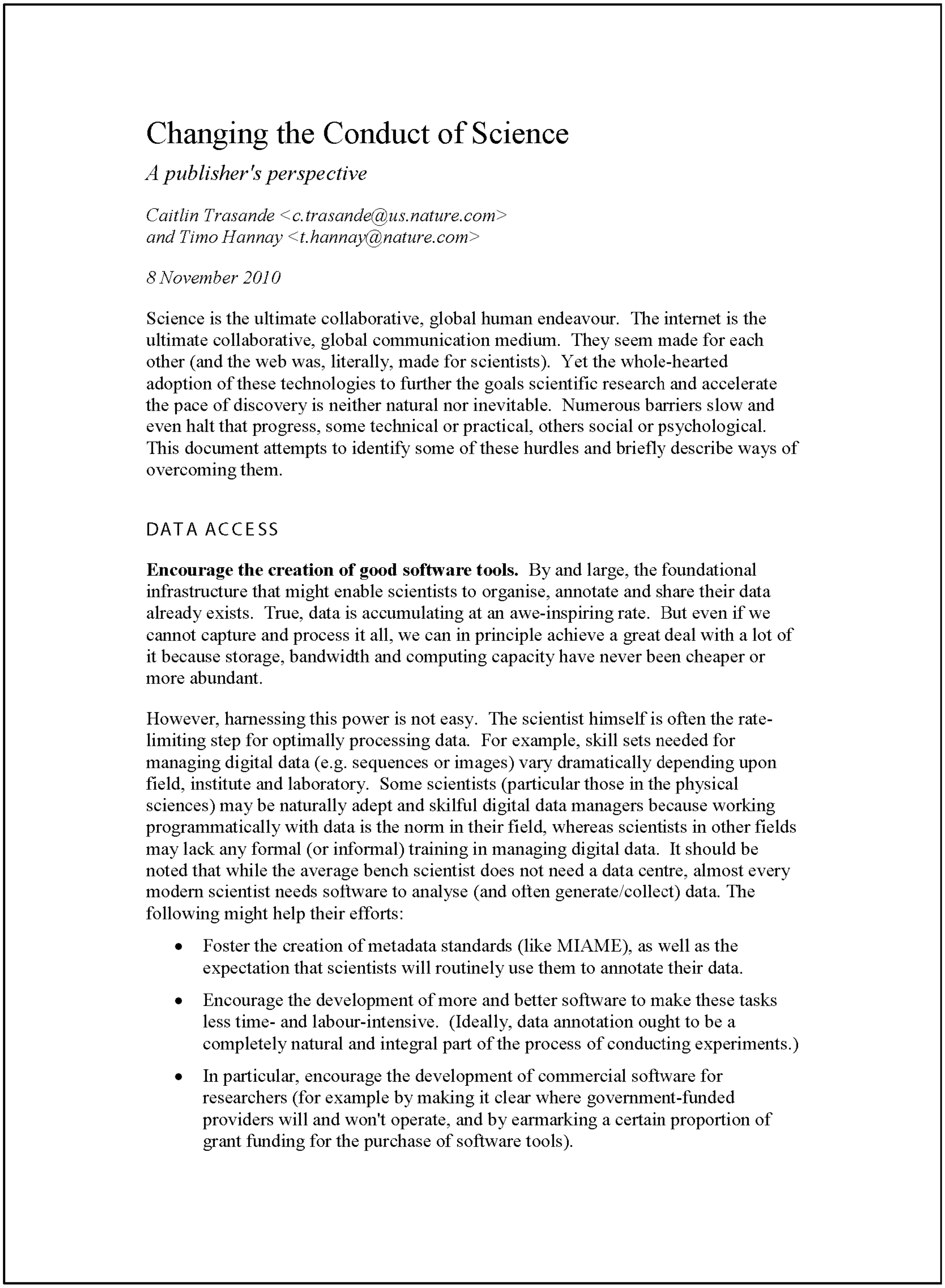
TALKING POINTS

- Social issues need clarity if the conduct of science is to change in a positive way:
  - Ownership of research data: *resolving the tension between confidentiality and openness; who decides?*
  - Intellectual property rights for open data: *attribution, publication policy on new results extracted from the data*
  - Rewards: *is the effort put into making data publicly accessible regarded as having the same kind of merit as the research that leads to a paper from the data?*
  - Is some data open by default? *Example, data that immediately supports the claims in a publication*
  - Policy issues for data repositories: *access to data, access to metadata, preservation policy, data integrity, replication, etc*
  - Standards: *Adoption/creation of international standards (metadata, formats, glossaries, indices) to facilitate search, retrieval, exploitation*
- Technical issues must be solved, requires support by research funders:
  - Access control and security for archives; protection of confidential data, facilitating automated access
  - Data integrity, preservation of provenance/attribution
  - Long-term preservation of data and access to it
  - Metadata (the richer the better!): distribution, creation of indexes, creation of federated data sets
  - Bandwidth and replication: the physical location of data when it is used
  - Local or remote computing (Grid, Cloud): the physical location of the computing resources that use the data
  - Creation of open community-usable tools to handle data, interface with local computing, perform intelligent search and retrieval
  - User features: ability to "browse" data before extracting it; local help desks; linking repositories with virtual research labs; ….

B F Schutz
Albert Einstein Institute

Talking points on data access || NSF: Changing the Conduct of Science in the Information Age || 1

Trasande, Caitlin, and Timo Hannay, "Changing the Conduct of Science: A Publisher's Perspective"



Changing the Conduct of Science

A publisher's perspective

Caitlin Trasande <c.trasande@us.nature.com>
and Timo Hannay <t.hannay@nature.com>

8 November 2010

Science is the ultimate collaborative, global human endeavour. The internet is the ultimate collaborative, global communication medium. They seem made for each other (and the web was, literally, made for scientists). Yet the whole-hearted adoption of these technologies to further the goals scientific research and accelerate the pace of discovery is neither natural nor inevitable. Numerous barriers slow and even halt that progress, some technical or practical, others social or psychological. This document attempts to identify some of these hurdles and briefly describe ways of overcoming them.

DATA ACCESS

Encourage the creation of good software tools. By and large, the foundational infrastructure that might enable scientists to organise, annotate and share their data already exists. True, data is accumulating at an awe-inspiring rate. But even if we cannot capture and process it all, we can in principle achieve a great deal with a lot of it because storage, bandwidth and computing capacity have never been cheaper or more abundant.

However, harnessing this power is not easy. The scientist himself is often the rate-limiting step for optimally processing data. For example, skill sets needed for managing digital data (e.g. sequences or images) vary dramatically depending upon field, institute and laboratory. Some scientists (particular those in the physical sciences) may be naturally adept and skilful digital data managers because working programmatically with data is the norm in their field, whereas scientists in other fields may lack any formal (or informal) training in managing digital data. It should be noted that while the average bench scientist does not need a data centre, almost every modern scientist needs software to analyse (and often generate/collect) data. The following might help their efforts:

- Foster the creation of metadata standards (like MIAME), as well as the expectation that scientists will routinely use them to annotate their data.
- Encourage the development of more and better software to make these tasks less time- and labour-intensive. (Ideally, data annotation ought to be a completely natural and integral part of the process of conducting experiments.)
- In particular, encourage the development of commercial software for researchers (for example by making it clear where government-funded providers will and won't operate, and by earmarking a certain proportion of grant funding for the purchase of software tools).

- Engage not just with publishing houses but also software houses and research-equipment suppliers.
- Give credit to those researchers who are genuinely open with their data, particularly when the data are used by others as a basis for their own research (see Attribution section below).
- Develop a consensus of what basic digital data management and computing skills are necessary to support high-quality data collection, processing, annotation and management.
- Fund training programs or the development of courses to ensure that all scientists have adequate competency in digital data management.

KNOWLEDGE ACCESS

Reward knowledge sharing of all kinds. Modern science was born when a reward structure was created that encouraged researchers to share their findings with each other through pages of academic journals rather then keeping them to themselves. Unfortunately these same incentives now have the perverse effect of discouraging knowledge sharing by other means. And as the opportunities for communication in the online world multiply (discussion forums, recommendations, wikis, file-sharing sites, blogs, microblogs, comments, votes, and so on), the aggregate cost of these lost opportunities grows. Whilst it is true that not all of these new means of communication are equally well suited to scholarship (and perhaps that some of them are downright counter-productive), the main reason that they are hardly exploited in research is the fact that contributions of these kinds are not tracked or rewarded. In the current incentive structure of science the author of the most influential academic blogs is trumped by the author of the most inconsequential peer-reviewed paper. This is patently wrong. Here are some ways it might be righted:

- Explicitly reward acts of self-arching in funder, institutional or other repositories by tracking this activity, making the statistics available, and using them in funding and appointment decisions.
- Similarly encourage the acts of posting preprints and blog entries, as well as commenting on them. (Systems for ranking these contributions by quality will be required, and any such system is vulnerable to gaming, but countermeasures are also possible so this is not an insurmountable challenge.)
- Conversely, recognise that in certain circumstances access restrictions are a feature, not a bug. (For example, where certain types of medical information are concerned, and where a truly open discussion can only take place away from the gaze of the globe and posterity.)
- Design a reward system for scientists who make themselves (and their reagents, algorithms, etc.) available to others. This could be piloted by tracking explicit acts of mentoring (e.g. evaluating PhD supervision).

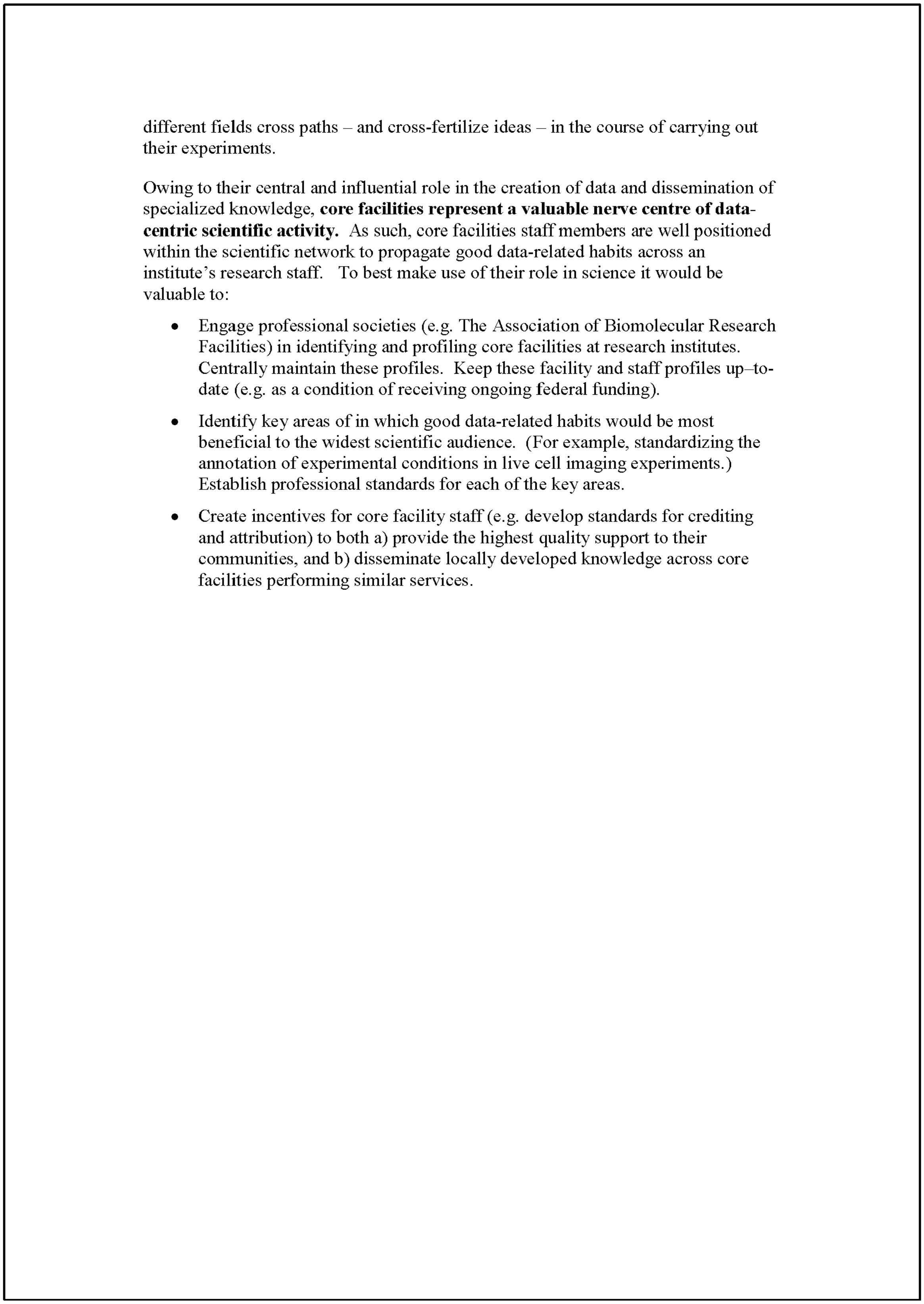
ATTRIBUTION

Support and use ID systems for researchers. It is a truth universally acknowledged that actions are driven by incentives, and incentives by attribution and credit. A substantial change in the way that science is conducted is difficult to imagine without a corresponding change in the way that academic credit is tracked and assigned. It is unfortunate, therefore, that perhaps the only scientifically significant objects in the known universe that lack a robust identification system are scientists themselves. Until the assignment and use of personal identifiers becomes routine, it will be next to impossible to track and reward the wide range of activities in which a 21st-century scientist ought to be engaged. Here are some ways of making it happen:

- Support and use of identifiers for researchers as a way of assigning credit for a wide variety of contributions, and making the decisions that stem from this. Encouraging the use of ORCIDs would be a good start: publishers should make their creation and capture an integral part of the editorial process; funders should use them to reward contributions to the common good.
- While it is unrealistic (and arguably undesirable) to aim for One True Identity System, a wild proliferation of systems would be counter-productive, so the creation of new ones where existing ones suffice should be avoided. Furthermore, interoperability with other identity systems is key – any system that does not readily interoperate with others does not deserve support.
- Encourage researchers to see their IDs more like loyalty cards (i.e., a means to gain credit for their contributions) than as social-security numbers (i.e., oppressive instruments of a potentially intrusive bureaucracy). Instill confidence that identities and related data are secure. In this regard, useful lessons might be learned from certain consumer markets.
- Recognise that identity (e.g., ORCID) is different from authentication (e.g.. OpenID). Though the two are related, they are best kept distinct and should not be confused with one another.
- Encourage the wide dissemination of activity data associated with personal IDs, and hence the creation of a wide range of derived metrics and rankings. Critics who point out that scientific research is too complicated to be measured are correct, which is precisely why we need a proliferation of metrics to encapsulate this complexity. This can only be provided by an open, competitive market for metrics.

DATA GENERATION: PLACES, PEOPLE AND TRAINING

Support hubs of scientific activity and training. Core facilities are institutionally managed shared experimentation resources (e.g. DNA and protein sequencing, light and electron microscopy, mass spectrometry). They are professionally staffed and designed to provide expert-led access to speciality equipment and technologies. Core facility directors and staff often provide an array of services, including training on specimen or sample preparation, operation of equipment and software, data collection and analysis, as well as experimental design and interpretation of experimental results. Core facilities represent a unique physical space where scientists from

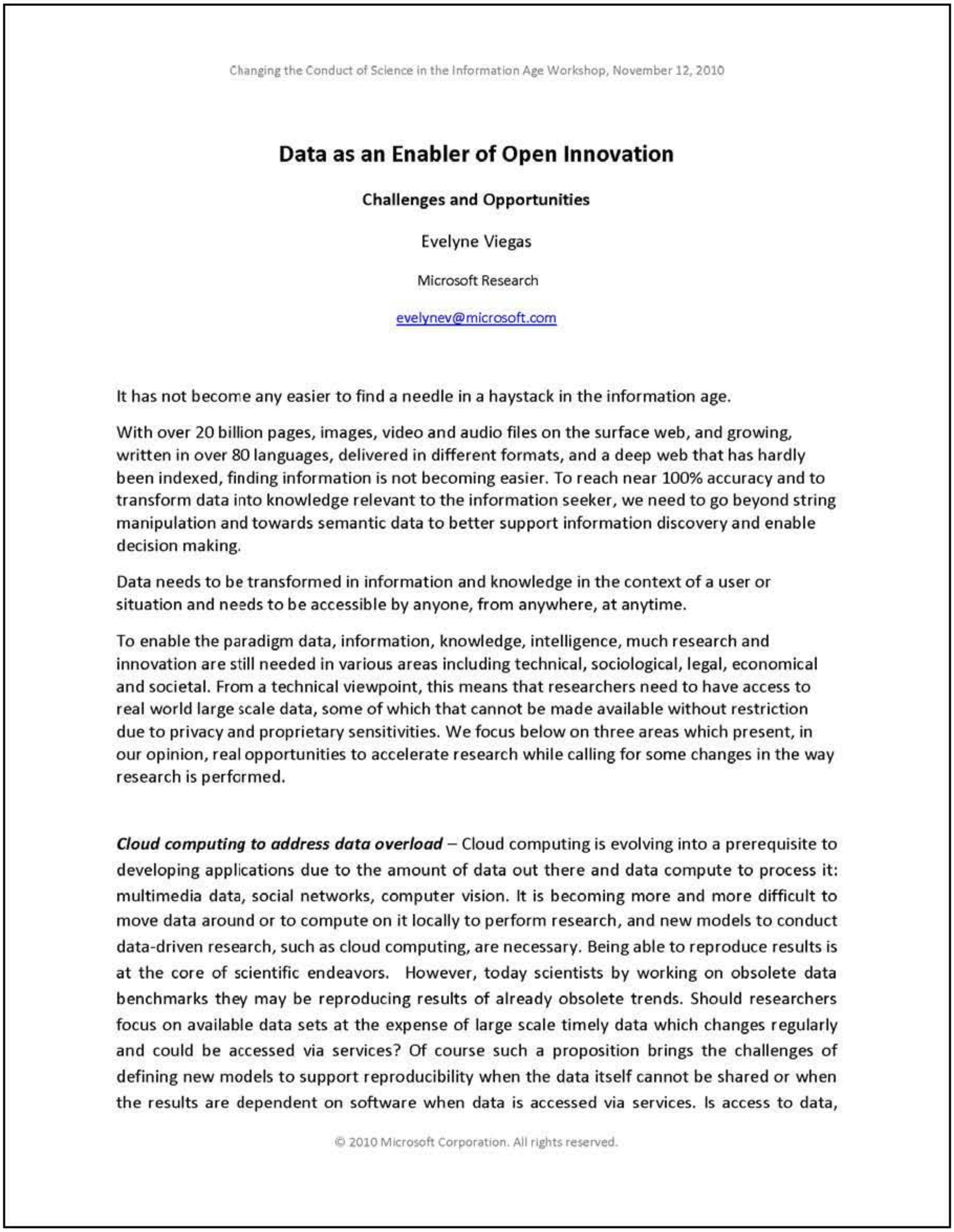


different fields cross paths – and cross-fertilize ideas – in the course of carrying out their experiments.

Owing to their central and influential role in the creation of data and dissemination of specialized knowledge, **core facilities represent a valuable nerve centre of data-centric scientific activity.** As such, core facilities staff members are well positioned within the scientific network to propagate good data-related habits across an institute's research staff. To best make use of their role in science it would be valuable to:

- Engage professional societies (e.g. The Association of Biomolecular Research Facilities) in identifying and profiling core facilities at research institutes. Centrally maintain these profiles. Keep these facility and staff profiles up–to-date (e.g. as a condition of receiving ongoing federal funding).
- Identify key areas of in which good data-related habits would be most beneficial to the widest scientific audience. (For example, standardizing the annotation of experimental conditions in live cell imaging experiments.) Establish professional standards for each of the key areas.
- Create incentives for core facility staff (e.g. develop standards for crediting and attribution) to both a) provide the highest quality support to their communities, and b) disseminate locally developed knowledge across core facilities performing similar services.

Viegas, Evelyne, "Data as an Enabler of Open Innovation: Challenges and Opportunities"



Data as an Enabler of Open Innovation

Challenges and Opportunities

Evelyne Viegas

Microsoft Research

evelynev@microsoft.com

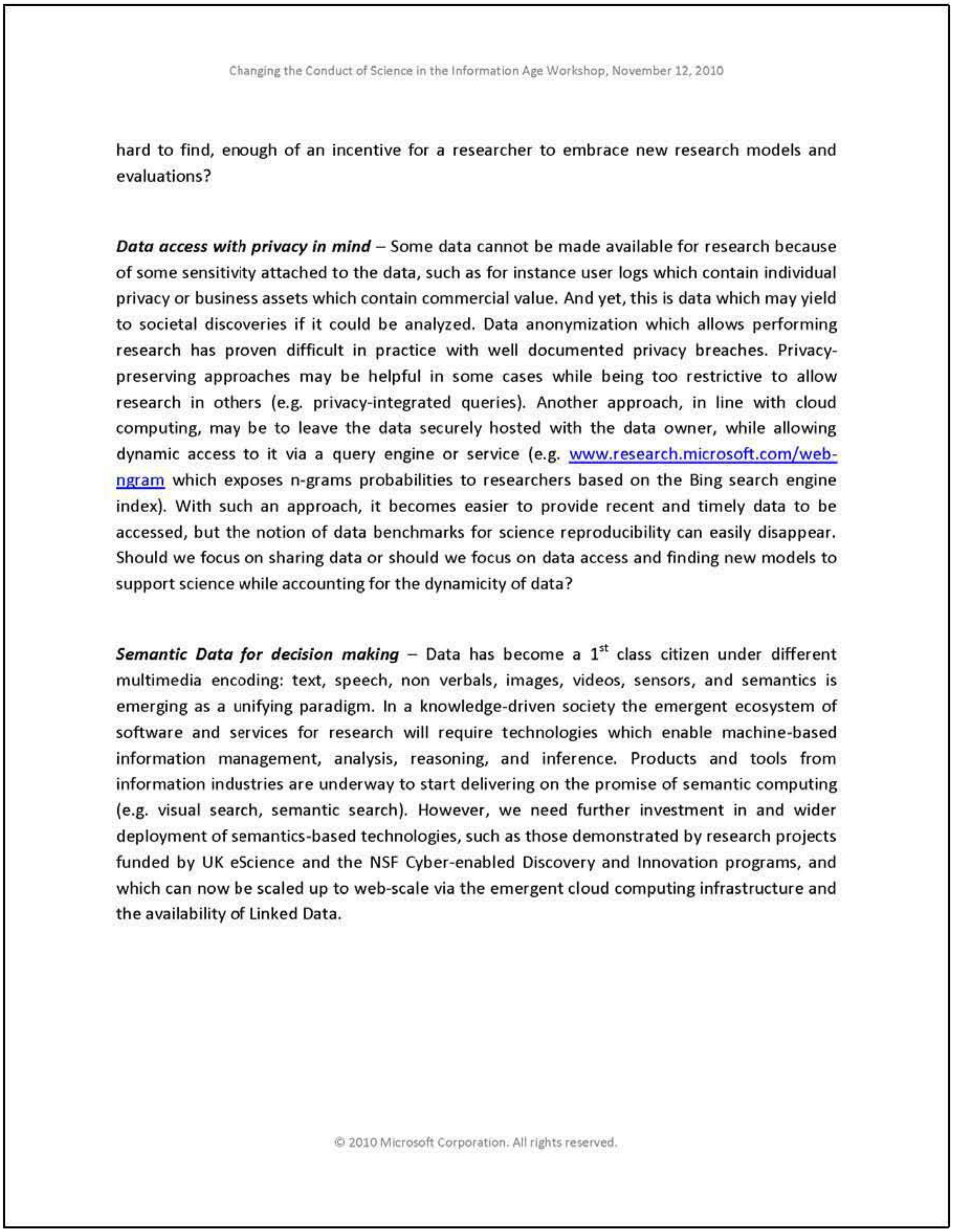
It has not become any easier to find a needle in a haystack in the information age.

With over 20 billion pages, images, video and audio files on the surface web, and growing, written in over 80 languages, delivered in different formats, and a deep web that has hardly been indexed, finding information is not becoming easier. To reach near 100% accuracy and to transform data into knowledge relevant to the information seeker, we need to go beyond string manipulation and towards semantic data to better support information discovery and enable decision making.

Data needs to be transformed in information and knowledge in the context of a user or situation and needs to be accessible by anyone, from anywhere, at anytime.

To enable the paradigm data, information, knowledge, intelligence, much research and innovation are still needed in various areas including technical, sociological, legal, economical and societal. From a technical viewpoint, this means that researchers need to have access to real world large scale data, some of which that cannot be made available without restriction due to privacy and proprietary sensitivities. We focus below on three areas which present, in our opinion, real opportunities to accelerate research while calling for some changes in the way research is performed.

Cloud computing to address data overload – Cloud computing is evolving into a prerequisite to developing applications due to the amount of data out there and data compute to process it: multimedia data, social networks, computer vision. It is becoming more and more difficult to move data around or to compute on it locally to perform research, and new models to conduct data-driven research, such as cloud computing, are necessary. Being able to reproduce results is at the core of scientific endeavors. However, today scientists by working on obsolete data benchmarks they may be reproducing results of already obsolete trends. Should researchers focus on available data sets at the expense of large scale timely data which changes regularly and could be accessed via services? Of course such a proposition brings the challenges of defining new models to support reproducibility when the data itself cannot be shared or when the results are dependent on software when data is accessed via services. Is access to data,



hard to find, enough of an incentive for a researcher to embrace new research models and evaluations?

Data access with privacy in mind – Some data cannot be made available for research because of some sensitivity attached to the data, such as for instance user logs which contain individual privacy or business assets which contain commercial value. And yet, this is data which may yield to societal discoveries if it could be analyzed. Data anonymization which allows performing research has proven difficult in practice with well documented privacy breaches. Privacy-preserving approaches may be helpful in some cases while being too restrictive to allow research in others (e.g. privacy-integrated queries). Another approach, in line with cloud computing, may be to leave the data securely hosted with the data owner, while allowing dynamic access to it via a query engine or service (e.g. www.research.microsoft.com/web-ngram which exposes n-grams probabilities to researchers based on the Bing search engine index). With such an approach, it becomes easier to provide recent and timely data to be accessed, but the notion of data benchmarks for science reproducibility can easily disappear. Should we focus on sharing data or should we focus on data access and finding new models to support science while accounting for the dynamicity of data?

Semantic Data for decision making – Data has become a 1st class citizen under different multimedia encoding: text, speech, non verbals, images, videos, sensors, and semantics is emerging as a unifying paradigm. In a knowledge-driven society the emergent ecosystem of software and services for research will require technologies which enable machine-based information management, analysis, reasoning, and inference. Products and tools from information industries are underway to start delivering on the promise of semantic computing (e.g. visual search, semantic search). However, we need further investment in and wider deployment of semantics-based technologies, such as those demonstrated by research projects funded by UK eScience and the NSF Cyber-enabled Discovery and Innovation programs, and which can now be scaled up to web-scale via the emergent cloud computing infrastructure and the availability of Linked Data.

References

Altman, Micah, and Gary King. 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine* 13 (3/4).

Aragão, Carlos. 2010. Lattes Platform. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Börner, Katy. 2010. Briefing Document. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Conlon, Michael. 2010. The objects of science and their representation in eScience. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Credit Where Credit Is Due. 2009. *Nature*, 16 December 2009.

Donoho, David. 2010. An Invitation to Reproducible Computational Research *Biostatistics* 11 (3):385-388.

Donoho, David, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. 2009. Reproducible Research in Computational Harmonic Analysis *Computing in Science and Engineering* 11 (8).

Elias, Peter. 2010. Digital Technology and the Conduct of Scientific Research. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

EU News Brief. 2010. European Union Observes Data Protection Day. http://eurunion.org/emailcampaigns/preview.php?previewtype=html&nl=21&c=237&m=158&s=6b9351a6f2b8c87a0a5d1e223a2907d4.

Evans, James A. 2010. Identification and the Complex System of Research. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Fenner, Martin. 2010. Personal Communication.

———. 2010. White Paper for "Changing the Conduct of Science in the Information Age". In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

German Data Forum (RatSWD). 2010. Recommendations for Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. In *Changing the Conduct of Science in the Information Age* National Science Foundation, Arlington, Virginia

Hirsh, Haym. 2010. How do you Cite a Crowd? In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Lambe, Patrick. 2010. Changing the Conduct of Science in the Information Age--Discussion Points. In *Changing the Conduct of Science in the Information Age*. National Science Foundaiton, Arlington, Virginia.

Lane, Julia. 2010. Let's Make Science Metrics More Scientific. *Nature*, http://www.nature.com/nature/journal/v464/n7288/full/464488a.html.

Lauer, Gerhard. 2010. Focusing on Sharing Knowledge and Data. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

National Science Foundation. 2010a. Proposal and Award Policies and Procedures Guide January 2010.

———. 2010b. Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans. *Press Release 10-077*, May 10, 2010.

———. 2011. Grants.gov Application Guide: A Guide for Preparation and Submission of NSF Applications via Grants.gov.

National Science Foundation, Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges. 2010. Cyber Science and Engineering.

Neylon, Cameron. 2010a. Warning: Misusing the Journal Impact Factor Can Damage Your Science! In *Science in the Open: The Online Home of Cameron Neylon*.

———. 2010b. Attribution and Identity for Researchers and Research Objects. In *Changing the Conduct of Science in the Informaiton Age*. National Science Foundation, Arlington, Virginia.

Office of Science and Technology Policy. 2009. Harnessing the Power of Digital Data for Science and Society.

Pfeiffenberger, Hans. 2010. Focusing on Social Constructs. In *Changing the Conduct of Science in the Information Age*. National Science Foundaiton, Arlington, Virginia.

Raddick, M., and A. Szalay. 2010. The Universe Online. *Science* (5995), http://www.sciencemag.org/content/329/5995/1028.full.

Schutz, Bernard F. 2010. Data Access: Digital Technology and Scientific Communities Talking Points. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Seidel, Edward. 2010. Data-intensive Transformation of Modern Science. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Stodden, Victoria. 2008. Enabling Reproducible Research: Open Licensing for Scientific Innovation. In http://www.stanford.edu/~vcs/papers/Licensing08292008.pdf.

———. 2010. Data Access: Digital Technology and Multiple Scientific Communities. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

———. 2010. Open Science: Policy Implications for the Evolving Phenomenon of User-Led Scientific Innovation. *Journal of Science Communication* 9 (1).

Trasande, Caitlin, and Timo Hannay. 2010. Changing the Conduct of Science: A publisher's perspective. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Viegas, Evelyne. 2010. Data as an Enabler of Open Innovation: Challenges and Opportunities. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Wood, John. 2010. A Vision for European Research 2030. In *Changing the Conduct of Science in the Information Age*. National Science Foundation, Arlington, Virginia.

Yale Law School Roundtable on Data and Code Sharing. 2010. Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science. *Computing in Science and Engineering* September/October 2010:8-12.