


## 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, 01-Mar-2006 11:52:46 EST by: Mark B. Orens

➤ Demonstrates the Power of the Web

## Webometrics Google Scholar Ranking (July 2010)

1	Harvard	
2	MIT	
3	UNAM	Southampton # 21
4	Minnesota	VirginiaTech # 37
5	UC Madrid	Cambridge # 97
6	Munich	Oxford # 115
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	

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

Home | Digital Curation | About Us | News | Events | Resources | Training | Projects | Community | Contact Us

**Digital Preservation Training Programme**  
London, 4 - 6 October 2010

**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.

**News Events**

- DCC UNLOCKS OPEN SCIENCE**  
18 September, 2010 | in DCC News
- Supporting Research Data Management at ECOL 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

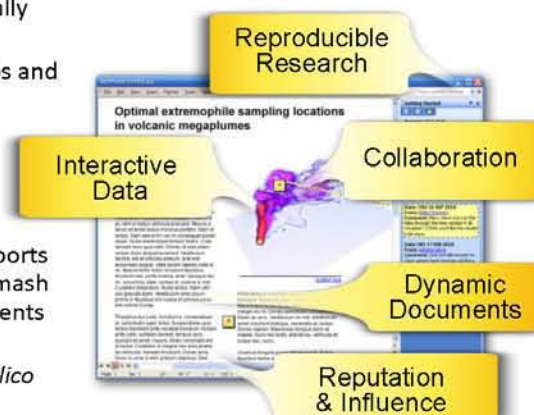
In his last talk Jim Gray highlighted three key areas for action relating to the future of Scholarly Communication and Libraries:

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

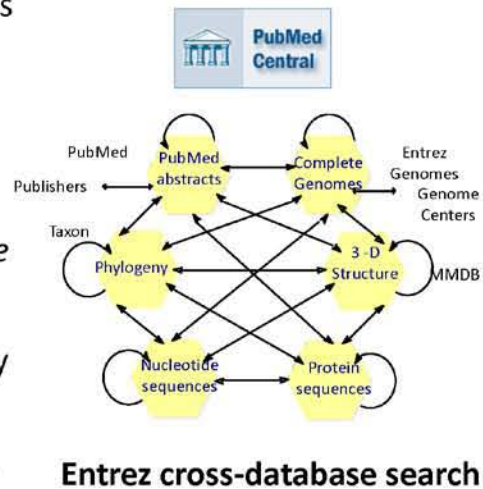


## Future of Research Libraries?

- 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.



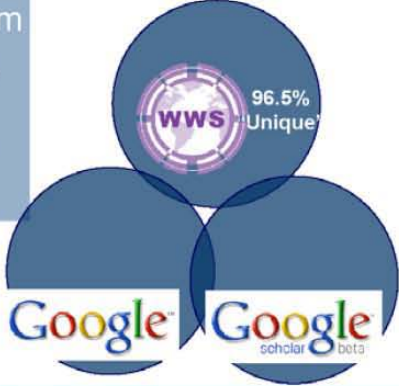
## WorldWideScience – Facts and Figures

- 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

A world map with a grid overlay, where the landmasses are shaded in a light purple color. The map is centered on the Atlantic Ocean, showing the Americas on the left and Europe, Africa, and Asia on the right.

## WorldWideScience – Fills Key Niche in Scientific Discovery

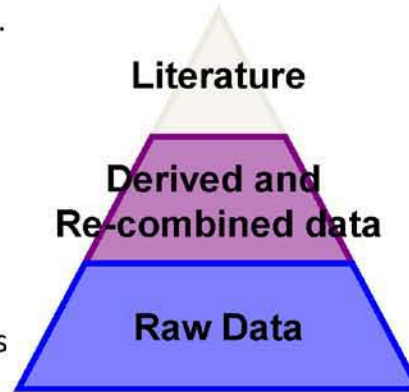
- In comparison of search results from identical queries on WWS, Google, and Google Scholar, only 3.5% overlap (i.e., WorldWideScience is 96.5% unique)

A Venn diagram consisting of three overlapping circles. The top circle is blue and contains the text 'WWS' and '96.5% Unique'. The bottom-left circle is blue and contains the text 'Google'. The bottom-right circle is blue and contains the text 'Google scholar beta'. The circles overlap in the center and at the intersections between two circles.

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



Slide from Jim Gray's last talk

