The Road to Open Access

Tony Hey
Vice President
Microsoft Research
Outline

• Some Open Access Pre-History
• A Journey to Open Access
• Research Repositories
• The Data Deluge and the Fourth Paradigm
• Open Data and Open Science
• An Open Access Future
Three Prophets of Open Access

• Paul Ginsparg
  • Creator of arXiv, an open access repository for pre-publication of much of the physics and astronomy literature

• David Lipmann
  • Pioneer at NCBI of the US National Library of Medicine’s PubMed Central Open Access publication repository and PubChem

• Stevan Harnad
  • Author of the subversive proposal, founder of Cogprints and tireless evangelist for Green Open Access
Harnad’s June 1994 Subversive Proposal

• The "Subversive Proposal" was an Internet posting by Stevan Harnad on June 27 1994 (presented orally at the 1994 Network Services Conference in London on November 28 1994)
• Called on all authors of "esoteric" writings—written only for research impact, not for royalty income—to archive them free for all online
• This led to the creation in 1997 of Cogprints, an Open Access Archive for self-archived articles in the Cognitive Sciences The Eprints software for creating OAI-compliant Open Access Institutional Repositories also evolved out of the Subversive Proposal.

May 1995 – JISC ELiB
Open Journals Project

A project that built on the concepts of open hypermedia (as exemplified by previous Microcosm and DLS projects) to create a framework for publishing that allowed journals on the Web to be interlinked and that would increase the readers' ability to follow, search and access the literature. The project developed open linking and citation linking framework, and led to the OpCit project.
Jan 1996 – JISC ELiB Cogprints Project

An ELiB Phase 2 program in “Preprints and the grey literature”, that aimed to produce an analogue to arxiv.org, a subject repository for the cognitive sciences. Project director was Stevan Harnad, then from the Department of Psychology at the University of Southampton.
Oct 1999 – EPrints announced

The first activity of the OpCit project was to fund Stevan Harnad and Les Carr to attend the UPS meeting in Santa Fe, subsequently renamed the Open Archiving Initiative (OAI). At the end of the two-day meeting that defined OAI-PMH, Stevan Harnad announced that Southampton would develop a turnkey system to help groups start up their own archives: this became EPrints.
The project, funded by the Joint NSF - JISC International Digital Libraries Research Programme, brought together Stevan Harnad (who had recently joined ECS) with Les Carr and Steve Hitchcock to investigate citation linking in the online literature.

Scale: to hyperlink each of the over 100,000 papers in Los Alamos's unique online Physics Archive to every other paper in the archive that it cites.

Compatibility: to develop and integrate a family of generic linking tools and to design author and user interfaces to enable easy adoption by other archives.

Universality: to promote the power of this remarkable new way of navigating the scientific journal literature and induce authors in other fields to create interlinked online archives like Los Alamos across disciplines and around the world.
Jan 2000 – First EPrints repository goes live (eprints.ecs.soton.ac.uk)

Jun 2000 – EPrints Software v1.0 released at the first official OAI workshop

Jan 2001 – Cogprints converts to EPrints
Dec 2001 – Budapest Open Access Initiative (BOAI)

The Budapest Open Access Initiative arises from a small but lively meeting convened in Budapest by the Open Society Institute (OSI) on December 1-2, 2001. The purpose of the meeting was to accelerate progress in the international effort to make research articles in all academic fields freely available on the internet.
Jan 2002 – Celestial OAI-PMH proxy harvesting service

An output of the OpCit project, Celestial aims to offer a comprehensive and accurate cache of up-to-date records from known OAI repositories. If OAI service providers harvest from this service, the load on data providers will be reduced. An effective aggregator will improve the interoperability, scalability and reliability of OAI services.
Jan 2002 – Citebase
Building on OpCit's bibliometric analysis work citation database, the Celestial service provides a citation-ranked view of the Open Access literature.

Feb 2002 – EPrints v2.0 released
Feb 2002 – International meeting on National Policies on Open Access

An invitation-only international meeting, joined by a major group of Pro-VCs and institutional digital information managers from Australia
Aug 2002 – Southampton Institutional Repository

Developed by the TARDis project, the Institutional Repository was a novel departure from the mainly discipline-specific repositories of the pre-OAI world.

TARDis Project Team:
Pauline Simpson and Jessie Hey
It is our policy to maximise the visibility, usage and impact of our research output by maximising online access to it for all would-be users and researchers worldwide. It is also our policy to minimise the effort that each of us has to expend in order to provide open online access to our research output.
The purpose of ROARMAP is to register and record the open-access policies of those institutions who are putting the principle of Open Access (as expressed by the Budapest Open Access Initiative and the Berlin Declaration) into practice. Currently lists 211 open access mandates internationally.
Jan 2004 – Registry of Open Access Repositories (ROAR)

Created initially to track the growth of EPrints repositories, and quickly expanded to deal with repositories based on any software.
ChronoZoom – History in its broadest possible context

The challenge: exploration of all known time series, and smoothly transition from billions of years down to individual nanoseconds...

This is what Walter Alvarez, Professor of Earth and Planetary Science at University of Berkeley set out to do. And he did it, with the help of Microsoft Research and the Live Labs team.

Our vision is to create an application that allows researchers to browse, overlay, and explore interdisciplinary data sources.

www.chronozoomtimescale.org

Open Access Timeline
http://test.chronozoomproject.org/southampton/openaccess/
Outline

• Some Open Access Pre-History
• A Journey to Open Access
• Research Repositories
• The Data Deluge and the Fourth Paradigm
• Open Data and Open Science
• An Open Access Future
Paul Ginsparg and arXiv
The Inevitability of Research 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

• nor to purchase conference proceedings and workshop contributions

➢ Insisted on keeping a digital copy of all research output in a Repository ...
Library Budgets and Journal Prices

MIT Libraries Materials Purchases vs. CPI % Increase
1986-2006

Percentage Change

-50% -0% 50% 100% 150% 200% 250% 300% 350% 400%

Year


- Consumer Price Index % +
- Serial Expenditures % +
- # Serials Purchased % +
- # Books Purchased % +
- Book Expenditures % +
Stevan Harnad and his 1994 Subversive Proposal
Green Open Access

‘Green Open Access’ or ‘Self-Archiving’ requires authors to make peer-reviewed final drafts of their articles accessible by depositing them in their Institution's OA Repository

• either on submission
• or on acceptance for publication

Notes:
• Individual papers need not be set to be immediately visible outside the institution - can be set to ‘delayed open access’ as in NLM’s PubMedCentral.
• Web copies of non-journal versions are allowed by over 80% of publishers ...
Leslie Carr and the EPrints Repository Software
## Flexible Repository Software

EPrints supports research papers, theses, teaching materials, arts and more.

## Getting Started

Find out more about the [EPrints software](#) or download it now:

**Latest Version:** v3.3.11
- [tar.gz](#)
- [deb](#)
- [rpm](#)

- **Debian/Ubuntu**
  - [Install Guide](#)
- **Redhat/Fedora**
  - [Install Guide](#)
- **Windows XP/Vista/7**
  - [Install Guide](#)

### Older Versions

- 3.2.0
- 3.1.3
- 3.0.5

[Try Live Demo](#)

## Using EPrints

- **EPrints Services**
  - Support, Hosting, Training and Consultancy
- **Documentation**
  - All latest documentation can be found on the EPrints wiki
- **Technical Discussion**
  - [Subscribe](#) | [View Archive](#)
- **Announcements**
  - [Subscribe](#)
- **Training Materials**
  - All training materials from our workshops

---

**UCL Discovery**

Welcome to UCL Discovery - home to UCL Research. Use this site to explore the unique scale and diversity of UCL research and our global expertise. UCL authors can access RPS from this site, and use au ...

[More ...](#)
Southampton EPrints Services

Professional Digital Repository Solutions

We offer a complete range of services to support EPrints, including hosting, training and consultancy.

Why use EPrints Services?

- Ten years of international experience in repository development and deployment
- Service provided by the same team that develops EPrints
- Fast and efficient service
- Clear choice of service options
- Minimizes impact on staffing resources
- Sustainability of ongoing support capability
- Based at a leading research institution recognised for leadership in Open Access

Types of Service Offered

We can help you by:

- Hosting and maintaining your repository
- Customising the repository to your specifications
- Importing your legacy data
- Providing ongoing technical support
- Providing customised training for repository managers and IT personnel
- Advising on policy development
- Assisting with advocacy and promotion
- Providing expert project management from project proposal to product launch

Click here to see the packages and make an enquiry.
Repository facilitated eprint request button to provide for researcher needs during embargo

Request a copy


Plain Text (embargoed articles) - Repository staff only until 27 July 2008
40Kb

Email address
Enter your email address.
myemail@wherever.edu

Reason
You may enter a rationale for requesting this document.
Please send me a copy for research purposes

Request a copy
Harold Varmus: PMC and PLoS
Once posted to PubMed Central, results of NIH-funded research become more prominent, integrated and accessible, making it easier for all scientists to pursue NIH’s research priority areas competitively.

PubMed Central materials are integrated with large NIH research data bases such as Genbank and PubChem, which helps accelerate scientific discovery.

Clinicians, patients, educators, and students can better reap the benefits of papers arising from NIH funding by accessing them on PubMed Central at no charge.

Finally, the Policy allows NIH to monitor, mine, and develop its portfolio of taxpayer funded research more effectively, and archive its results in perpetuity.”
PMC Compliance Rate
- Before legal mandate compliance was 19%
- After legal mandate compliance up to 75%
- Signed into law by George W. Bush in 2007

NIH have taken a further step of announcing that, ‘sometime in 2013’ they ‘... will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.’
US Fair Access to Science and Technology Research (FASTR) Act

• “The United States has a substantial interest in maximizing the impact of the research it funds by enabling a wide range of reuses of the peer-reviewed literature reporting the results of such research, including by enabling automated analysis by state-of-the-art technologies.”

• Federal agencies consider whether or not the terms of use should include “a royalty free copyright license that is available to the public and that permits the reuse of those research papers, on the condition that attribution is given to the author or authors of the research and any others designated by the copyright owner”

13 February 2013
US White House Memorandum (1)

- Directive requiring the major Federal Funding agencies “to develop a plan to support increased public access to the results of research funded by the Federal Government.”
- The memorandum defines digital data “as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.”

22 February 2013
“Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.”
Outline

• Some Open Access Pre-History
• A Journey to Open Access
• Research Repositories
• The Data Deluge and the Fourth Paradigm
• Open Data and Open Science
• An Open Access Future
Future of Research Repositories?

- 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
- Need for federated databases of scientific information and cross database search tools
  - Centralized: NIH National Library of Medicine
  - Distributed: WorldWideScience.org
The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.

- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication.
- 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
Brief History: Federated Search and WorldWideScience.org

The Invisible Web

- Science is hundreds of times larger than the visible “Surface Web”
- “Invisible Web” generally not searchable by major search engines
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
Science Open Access Journals (SOAJ) is a gateway to open access publications in the sciences. Using federated search technology, SOAJ allows the user to search multiple open access publications with a single query. Results are consolidated and returned in relevance-ranked order, and users are directed to the originating sources for access to the full-text documents.
University Reputation Management

• Organizations like Webometrics will be using the university’s visible web presence to produce global and national rankings of universities
  ➢ Many different ranking methodologies, need to ensure university displays its best data.

• With the advent of open access to both full text of papers and data, university research repositories will be an important part of the university’s reputation management strategy
  ➢ COAR – Coalition of Open Access Repositories
COAR, the Confederation of Open Access Repositories, is a young association of repository initiatives launched in October 2009, uniting 59 institutions in 23 countries from throughout Europe, Latin America, Asia, and North America. Its mission is to enhance greater visibility and application of research outputs through global networks of Open Access digital repositories.

COAR promotes infrastructure interoperability and a joint global data store of Open Access repositories to enable and support the re-use of data by service and portal providers. Currently, COAR has three working groups, each with its own set of responsibilities, objectives, and related activities.
Webometrics University Rankings
http://www.webometrics.info/

Webometrics overall rankings are obtained by combining four different rankings with the following weights:

- Visibility (external inlinks) 50%
- Size (Web Pages) 20%
- Rich Files (pdf, doc, ps, ppt) 15%
- Scholar (papers and citations) 15%
## Webometrics: Top 25 Universities Worldwide

### July 2010

<table>
<thead>
<tr>
<th>#</th>
<th>DOMAINS</th>
<th>COUNTRY</th>
<th>REGION</th>
<th>SIZE</th>
<th>VISIBILITY</th>
<th>RICH</th>
<th>SCHOLAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>harvard.edu</td>
<td>us</td>
<td>North America</td>
<td>3</td>
<td>3</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>mit.edu</td>
<td>us</td>
<td>North America</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>stanford.edu</td>
<td>us</td>
<td>North America</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>berkeley.edu</td>
<td>us</td>
<td>North America</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>cornell.edu</td>
<td>us</td>
<td>North America</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>umich.edu</td>
<td>us</td>
<td>North America</td>
<td>8</td>
<td>7</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>umn.edu</td>
<td>us</td>
<td>North America</td>
<td>10</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>washington.edu</td>
<td>us</td>
<td>North America</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td>10</td>
<td>wisc.edu</td>
<td>us</td>
<td>North America</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>48</td>
</tr>
<tr>
<td>11</td>
<td>utexas.edu</td>
<td>us</td>
<td>North America</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>upenn.edu</td>
<td>us</td>
<td>North America</td>
<td>19</td>
<td>9</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>psu.edu</td>
<td>us</td>
<td>North America</td>
<td>1</td>
<td>45</td>
<td>5</td>
<td>116</td>
</tr>
<tr>
<td>14</td>
<td>columbia.edu</td>
<td>us</td>
<td>North America</td>
<td>14</td>
<td>14</td>
<td>19</td>
<td>89</td>
</tr>
<tr>
<td>15</td>
<td>cmu.edu</td>
<td>us</td>
<td>North America</td>
<td>11</td>
<td>27</td>
<td>3</td>
<td>74</td>
</tr>
<tr>
<td>16</td>
<td>uiuc.edu</td>
<td>us</td>
<td>North America</td>
<td>42</td>
<td>12</td>
<td>20</td>
<td>75</td>
</tr>
<tr>
<td>17</td>
<td>ucla.edu</td>
<td>us</td>
<td>North America</td>
<td>18</td>
<td>15</td>
<td>21</td>
<td>78</td>
</tr>
<tr>
<td>18</td>
<td>tamu.edu</td>
<td>us</td>
<td>North America</td>
<td>24</td>
<td>31</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>19</td>
<td>umd.edu</td>
<td>us</td>
<td>North America</td>
<td>25</td>
<td>28</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>20</td>
<td>purdue.edu</td>
<td>us</td>
<td>North America</td>
<td>33</td>
<td>32</td>
<td>14</td>
<td>39</td>
</tr>
<tr>
<td>21</td>
<td>unc.edu</td>
<td>us</td>
<td>North America</td>
<td>22</td>
<td>19</td>
<td>22</td>
<td>148</td>
</tr>
<tr>
<td>22</td>
<td>msu.edu</td>
<td>us</td>
<td>North America</td>
<td>21</td>
<td>24</td>
<td>26</td>
<td>100</td>
</tr>
<tr>
<td>23</td>
<td>cam.ac.uk</td>
<td>uk</td>
<td>Europe</td>
<td>15</td>
<td>18</td>
<td>54</td>
<td>97</td>
</tr>
<tr>
<td>24</td>
<td>ufl.edu</td>
<td>us</td>
<td>North America</td>
<td>23</td>
<td>37</td>
<td>12</td>
<td>94</td>
</tr>
<tr>
<td>25</td>
<td>rutgers.edu</td>
<td>us</td>
<td>North America</td>
<td>17</td>
<td>38</td>
<td>13</td>
<td>127</td>
</tr>
<tr>
<td>Rank</td>
<td>Institution</td>
<td>Position</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------------------</td>
<td>----------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Harvard</td>
<td>#1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MIT</td>
<td>#2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>UNAM</td>
<td>#3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Minnesota</td>
<td>#4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>UC Madrid</td>
<td>#5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Munich</td>
<td>#6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Stanford</td>
<td>#7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>U Queensland</td>
<td>#8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kyoto</td>
<td>#9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Masaryk</td>
<td>#10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Toronto</td>
<td>#11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Michigan</td>
<td>#12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>UPC Barcelona</td>
<td>#13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Texas A&amp;M</td>
<td>#14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ETH Zurich</td>
<td>#15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Nebraska</td>
<td>#16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Groningen</td>
<td>#17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Vienna</td>
<td>#18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>CUHK</td>
<td>#19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Georgia Tech</td>
<td>#20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Southampton</td>
<td>#21</td>
<td>(85)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Cornell</td>
<td>#22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pennsylvania</td>
<td>#23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Tokyo</td>
<td>#24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Murcia</td>
<td>#25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Worldwide Percentage Open Access (OA) by Field (2005-10)

**Green OA:** 26.3%

**Gold OA:** 4.3%

**Delayed Access:** 6%

**Total:** 36.6%
Green OA Mandates can triple Green OA

N= 63 518

Percentage

Non-Mandated Institutions  |  Mandated Institutions

2002  |  2003  |  2004  |  2005  |  2006  |  2007  |  2008  |  2009

0%  |  10%  |  20%  |  30%  |  40%  |  50%  |  60%  |  70%  |  80%
Deposit rate depends on mandate strength

*Mandate Strength:*

12 immediate deposit required + linked to performance evaluation (Liege) (no waiver option)
9 immediate deposit required (no waiver option)
6 6-month delay allowed (no waiver option)
3 12-month delay allowed (no waiver option)
3 rights-retention with waiver option (Harvard, MIT)
1 deposit if/when publisher says it's ok
0 no OA policy at all

**Mandate Strength** is positively correlated with **Total Deposit Count**

**Mandate Age** is positively correlated with **Total Deposit Count**

<table>
<thead>
<tr>
<th>Articles published 2010-2011 (ISI)</th>
<th>Harvard</th>
<th>MIT</th>
<th>U Liège</th>
</tr>
</thead>
<tbody>
<tr>
<td>27,588</td>
<td>8,247</td>
<td>2,734</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent Green on web</th>
<th>Harvard</th>
<th>MIT</th>
<th>U Liège</th>
</tr>
</thead>
<tbody>
<tr>
<td>62%</td>
<td>68%</td>
<td>49%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent deposited in Repository</th>
<th>Harvard</th>
<th>MIT</th>
<th>U Liège</th>
</tr>
</thead>
<tbody>
<tr>
<td>4%</td>
<td>29%</td>
<td>83%</td>
<td></td>
</tr>
</tbody>
</table>
Require immediate-deposit (whether green, gold or embargoed) of final peer-reviewed draft on date of acceptance in author's institutional repository monitored and verified by author's institution as funding compliance condition and institutional employment condition

Institutional repository designated mechanism for submitting publications for institutional REF performance evaluation, research grant applications and national research assessment

Repository deposits monitored to generate rich and visible metrics of usage and citation, to verify and reward authors for deposit and to showcase and archive the institution's and funder's research output and impact

Recommend (but do not require) immediate un-embargoed green or gold CC-BY
Outline

• Some Open Access Pre-History
• A Journey to Open Access
• Research Repositories
• The Data Deluge and the Fourth Paradigm
• Open Data and Open Science
• An Open Access Future
A Tidal Wave of Scientific Data
Emergence of a Fourth Research Paradigm

Thousand years ago – Experimental Science
• Description of natural phenomena

Last few hundred years – Theoretical Science
• Newton’s Laws, Maxwell’s Equations...

Last few decades – Computational Science
• Simulation of complex phenomena

Today – Data-Intensive Science
• Scientists overwhelmed with data sets from many different sources
  • Captured by instruments
  • Generated by simulations
  • Generated by sensor networks

\[
\left( \frac{a}{a} \right)^2 = \frac{4\pi G \rho}{3} - K \frac{c^2}{a^2}
\]

eScience is the set of tools and technologies to support data federation and collaboration
• For analysis and data mining
• For data visualization and exploration
• For scholarly communication and dissemination

(With thanks to Jim Gray)
Machine Learning and eScience

Tackling societal challenges

Identifying genetic and environmental causes of disease

Fighting HIV/AIDS

Increasing energy yield of sugar cane through genome assembly
The ‘Cosmic Genome’ Project

- The Sloan Digital Sky Survey is the first major astronomical survey project:
  - 5 color images and spectra of ¼ of the sky
  - Pictures of over 300 million celestial objects
  - Distances to the closest 1 million galaxies
- Jim Gray from Microsoft Research worked with astronomer Alex Szalay to build the public ‘SkyServer’ archive for the survey

> New model of scientific publishing - publish the data before astronomers publish their analysis
Outline

• Some Open Access Pre-History
• A Journey to Open Access
• Research Repositories
• The Data Deluge and the Fourth Paradigm
• Open Data and Open Science
• An Open Access Future
The Berlin Declaration 2003

• ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’

• Defines open access contributions as including:
  • ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’
Key Emerging Themes

Scholarship as we know it is changing...

- Funding Agencies promoting the Curation and Sharing of Data
- Increasing need for Research Reproducibility/Interoperability
- Changing the Reward System
NSF Data Sharing Policy 2010

“Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.”

All future grant proposals now require a two-page Data Management Plan that addresses the above requirement and the Plan will be subject to peer review.
The Government, in line with our overarching commitment to transparency and open data, is committed to ensuring that publicly-funded research should be accessible free of charge.

Free and open access to taxpayer-funded research offers significant social and economic benefits by spreading knowledge, raising the prestige of UK research and encouraging technology transfer.
Key driver from a UK Research Council

EPSRC Policy Framework on research data (May 2011):

• “all institutions in receipt of their funding should develop a clear roadmap for research data management, which should be implemented by May 1st 2015”

• “organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years”
Datacite and ORCID

DataCite

- International consortium to establish easier access to scientific research data
- Increase acceptance of research data as legitimate, citable contributions to the scientific record
- Support data archiving that will permit results to be verified and re-purposed for future study.

ORCID - Open Research & Contributor ID

- Aims to solve the author/contributor name ambiguity problem in scholarly communications
- Central registry of unique identifiers for individual researchers
- Open and transparent linking mechanism between ORCID and other current author ID schemes.
- Identifiers can be linked to the researcher’s output to enhance the scientific discovery process
A RESTful Interface for Data

Open Data Protocol

Just HTTP

- Items as resources, HTTP methods (GET, PUT, ...) to act
- Leverage proxies, authentication, ETags, ...

Uniform URL convention

- Every piece of information is addressable
- Predictable and flexible URL syntax

Multiple representations

- Use regular HTTP content-type negotiation
- JSON and Atom (full AtomPub support)

http://www.odata.org
Dataverse: Re-inforcing the Link between Research Publications and Research Data

The Dataverse project at Harvard has been awarded an Alfred Sloan Foundation grant for the next 2 years to enhance the link between journals and data.

Seamless integration between the two systems:

Deposit Data from Journal to Dataverse through a standard API (based on SWORD)
Outline

- Some Open Access Pre-History
- A Journey to Open Access
- Research Repositories
- The Data Deluge and the Fourth Paradigm
- Open Data and Open Science
- An Open Access Future
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
FACULTY of 1000

Faculty of 1000 is the publisher of four unique services that support and inform the work of life scientists and clinicians.

F1000Prime An in-depth directory of top articles in biology and medicine, as recommended by our Faculty of over 5,000 expert scientists and clinical researchers, assisted by 5,000 associates.

F1000Trials A comprehensive listing of published clinical trials with recommendations and comments on the most important. New from Faculty of 1000: beta-launching in Spring 2013.

F1000Research An innovative open access publishing program offering immediate publication and open peer review.

F1000Posters The open access repository for posters and slide presentations in biology and medicine.

Press Releases

Use of psychedelic drugs remains prevalent in the United States
23 Apr 2013
An article published in F1000Research, and approved for indexing in PubMed and other major bibliographical databases, estimates that there were approximately 32 million users of psychedelic drugs in the United States in 2010.
4.2 Polarisation Dependence

The polarisation dependence of the SHG signal was recorded for bulk phenylalanine concentrations from 0.005 mol dm$^{-3}$ up to 0.8 mol dm$^{-3}$. This approach would reveal any change in the preferred orientation of the phenylalanine molecule at the air/water interface with increasing bulk concentration. The SHG signal was monitored for the linear output polarisations $I_{P}$, $I_{S}$, $I_{PS}$, and $I_{P+S}$, where $P$, $S$, and $+45$ correspond to the output harmonic polarisation (Pharmonic and $\Gamma=90^\circ$ for Sharmonic light). The conditions were optimised to avoid the two photon fluorescence present with the SHG signal. A representative sample of five plots of data at different concentrations are shown in Figures 711.

Link to simulation software and data in archive

Link to data, follow links back to the raw data archive
• 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

(From Jim Gray’s last talk)
Future Research Infrastructure will use semantic knowledge services on Client + Cloud.
Have Librarians Abdicated?

- Commercial entities have stepped into the traditional library space—Yahoo, Google, (and yes) Microsoft ...
- Engineering students at Southampton rarely went to the library for ‘library purposes’ – wi-fi, coffee, meeting place ...
- Librarians are being disintermediated

"...Sleepwalking into the era of the iPhone, the eBook and the Xbox without a strategy," she suggested, "runs the risk of turning the library service into a curiosity of history such as telex machines or typewriters."

"If you don't like change, you're going to like irrelevance even less."

Margaret Hodge (Dec09)  
Britain’s Former Culture Minister  

General Eric Shinseki  
Retired United States Army four-star general, currently US Secretary of Veterans Affairs
A Possible Answer: The MIT DSpace Vision

“Much of the material produced by faculty, such as datasets, experimental results and rich media data as well as more conventional document-based material (e.g. articles and reports) is housed on an individual's hard drive or department Web server. Such material is often lost forever as faculty and departments change over time.”

An important role for Research Libraries is to be the guardians of the research output of the institution
Challenges for Research Libraries

• Major university research libraries in each discipline should take the lead in developing a federated repository system.
• Green repositories could easily be filled by Library leadership and assistance in self-archiving for high-profile university researchers.
• Many problems to be solved such as versioning, multiple authors - but that is the library’s job!
• Many challenges related to federated search and the ‘Invisible Web’
Commentary

This is one of a series of commentaries on the future of scientific publishing. For a listing of the other commentaries, see http://www.jneurosci.org/cgi/content/full/26/36/9077.

As We May Read

Paul Ginsparg
Departments of Physics and Information Science, Cornell University, Ithaca, New York 14853

The e-print arXiv (http://arXiv.org), initiated in August 1991, has effectively transformed the research communication infrastructure of multiple fields of physics and could play a prominent role in a unified set of global resources for physics, mathematics, and computer science. It has grown to contain >375,000 articles (as of July 2006), with >50,000 new submissions expected in calendar year 2006 and >40,000,000 full-text downloads per year. It is an international project, with dedicated mirror sites in 17 countries and orders of magnitude. Even with the majority of science research journals now on-line, researchers continue to enjoy both the benefits of the rapid availability of the materials, even if not yet reviewed, and open archival access to the same materials, even if held in parallel by conventional publishers. The methodology works within copyright law, as long as the depositor has the authority to deposit the materials and assign a nonexclusive license to distribute at the time of deposition, because such a license takes precedence over any subsequent ones.

helps ensure that the arXiv remains a forum for communication among research professionals, not a mechanism for outsiders to communicate to that community. Additionally, a small group of volunteer “moderators,” consisting of interested experts from around the world, cursorily precans new submissions, typically only at the level of title and abstract, for appropriateness to the proposed primary subject area.

The arXiv repository functions are flexible enough either to coexist with the
Ginsparg’s Conclusions?

“On the one-decade time scale, it is likely that more research communities will join some form of global unified archive system without the current partitioning and access restrictions familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge.”

“Ironically, it is also possible that the technology of the 21st century will allow the traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research Enterprise.”
Some Resources

- Microsoft Research
  - [http://research.microsoft.com](http://research.microsoft.com)
  - Microsoft Research downloads: [http://research.microsoft.com/research/downloads](http://research.microsoft.com/research/downloads)

- Microsoft Research Connections

- Science at Microsoft
  - [http://www.microsoft.com/science](http://www.microsoft.com/science)

- Scholarly Communications
  - [http://www.microsoft.com/scholarlycomm](http://www.microsoft.com/scholarlycomm)

- CodePlex
  - [http://www.codeplex.com](http://www.codeplex.com)

- Outercurve Foundation
  - [http://www.outercurve.org/](http://www.outercurve.org/)

- Tony Hey on eScience
  - [http://tonyhey.net/](http://tonyhey.net/)