Libraries in the Converging Worlds of Open Data, E-Research,

INFORMATION and communication technologies (ICT) are transforming the way academic researchers work. The new forms of research enabled by the latest technologies bring about collaboration among researchers in different locations, institutions, and even disciplines. These new collaborations have two key features—the prodigious use and production of data. This data-centric research manifests itself in such concepts as e-science, cyberinfrastructure, or e-research.

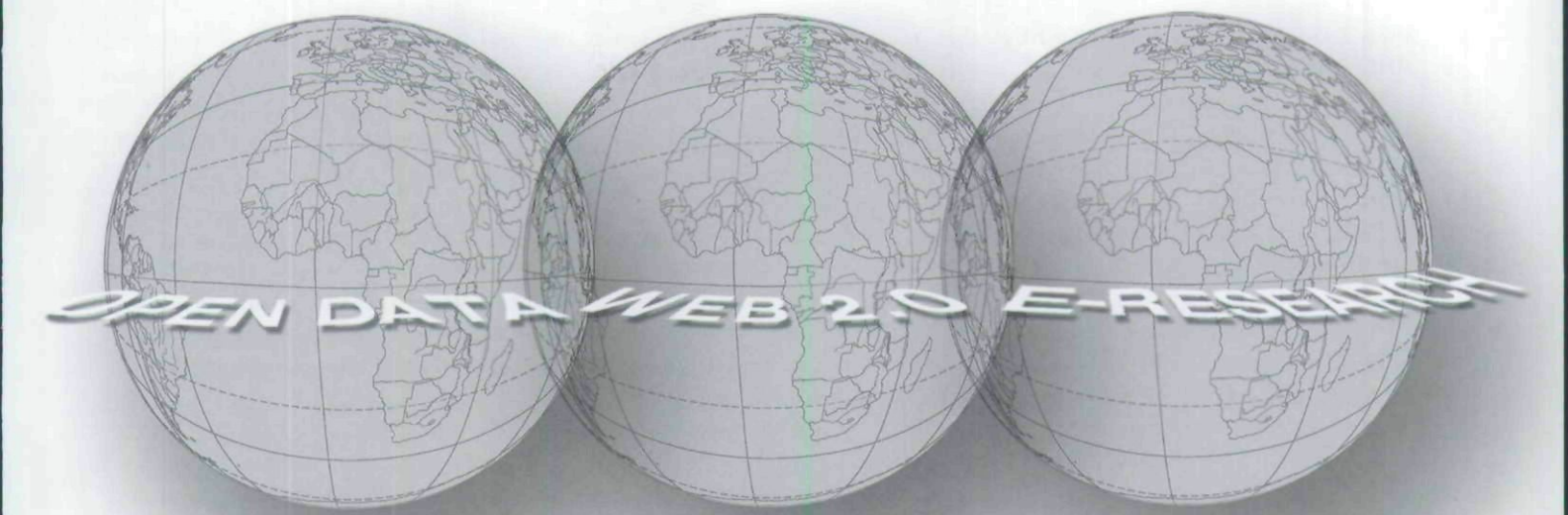
In order to make sense of the converging data-related initiatives, trends, and technologies, the DISC-UK (Data Information Specialists Committee) DataShare project aims to occupy a key position in bringing research libraries into the field of data curation, while supporting data management and e-research activities via open access institutional repositories and Web 2.0 technologies.

Recent research carried out by the Australian Department of Education, Science and Training ("Backing Australia's Ability—An Ongoing Commitment"; http://backingaus.innovation.gov.au/info_booklet/on_commit.htm) has indicated that the amount of data generated in the next 5 years will surpass the volume of data ever created. A recent IDC White Paper ("The Expanding Digital Universe—A Forecast of Worldwide Information Growth through 2010"; www.emc.com/about/destination/digital_universe) predicted that between 2006 and 2010, the information added annually to the digital universe will increase more than sixfold—from 161 exabytes to 988 exabytes. Such statements alone have significant implications for data storage, publishing, confidentiality, preservation, and curation. Indeed, should these predictions be accurate, the practice of managing such a data deluge will acquire a more prominent role within the research lifecycle. Researchers, librarians, technologists, publishers, and policymakers will have to adapt their practices in order to deal with this new landscape.

Traditionally, these actors played discrete roles in the research lifecycle from an initial concept to the eventual published output. The open access movement advocates the introduction of such players into a virtual space while streamlining the whole research lifecycle and making available the institutional research output in open environments for those who need it or want to access it.

OPEN DATA

Over the last decade there has been much discussion about the merits of open standards, open source software, open access to scholarly publications, and most recently open data. This discussion has tended to highlight that such initiatives would lead to institutional and community benefits in terms of greater accessibility to and long-term preservation of research output and of cost savings.



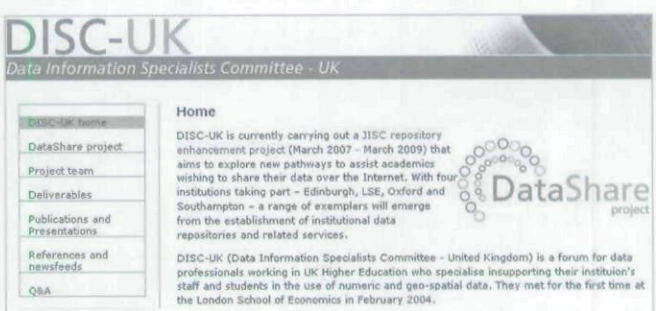
and Web 2.0

by Stuart MacDonald and Luis Martinez Uribe

The concept of open data was introduced in 2004 in the publication *OECD Principles and Guidelines for Access to Research Data from Public Funding* (www.oecd.org/data oecd/9/61/38500813.pdf).

Other august bodies, such as the National Science Foundation (NSF) in Chapter 3 of its *Cyberstructure Vision for 21st Century Discovery* (www.nsf.gov/od/oci/ci_v5.pdf), the Research Information Network (RIN)'s Data Principles (www.rin.ac.uk/data-principles), and the Joint Information Systems Committee (JISC) and the Office of Science and Innovation's *Developing the UK's e-infrastructure for science and innovation* (www.nesc.ac.uk/documents/OSI/report .pdf) have all contributed to the dialogue advocating open access to research data. Academic research is primarily based on positive data or results, but there are practitioners who believe that data from failed experimentation or "dark data" ("Freeing the Dark Data of Failed Scientific Experiments," by Thomas Goetz, *Wired*, September 2007; www.wired.com/ science/discoveries/magazine/115-10/st_essay), which constitutes the vast majority of data produced in academic research, has equal validity in terms of knowledge and as such should also be made freely available.

However, open data doesn't just happen by itself. Intimately linked to the altruism of the open data movement are technological, cultural, and legal issues that need to be addressed in order for what is now a global research community to reap the full benefits. Currently many researchers do not appear aware of or interested in issues surrounding their own data management. Some domains do have well-developed data curation strategies. However, as Liz Lyon reported in a recent JISC-funded Consultancy Report ("Dealing with Data"; www.ukoln.ac.uk/ukoln/staff/ e.j.lyon/reports/dealing_with_data_report-final.pdf), there is a real need for leadership and cross-domain thinking to effectively manage the data deluge. Higher education institutions need to take some responsibility with regard to implementing effective data management systems for research data outputs.



DISC-UK DataShare Project website (www.disc-UK.org/datashare.html)

WEB 2.0 FOR SOCIAL NETWORKING AND DATA PRESENTATION

The environment described above is complex, dynamic, and ever-changing. There are a number of resources embracing Web 2.0 technologies that aim to keep practitioners up-to-date with news and activities in this area.

There are a range of authoritative weblogs that address the open movement, some of which are included in the list below:

- The DCC's Digital Curation Blog (http://digitalcuration.blogspot.com)
- Peter Suber's Open Access News (www.earlham.edu/~peters/fos/fosblog.html)
- The Research Information Network's Team Blog (www.rin.ac.uk/team-blog)
- Open Knowledge Foundation Weblog (http://blog.okfn.org)
- Peter Murray Rust's Blog (http://wwmm.ch.cam.ac.uk/blogs/murrayrust)
- OA Librarian (http://oalibrarian.blogspot.com)

There are also a number of Facebook groups addressing the subject of open access:

- Librarians Who Support Open Access
- SPARC (Scholarly Publishing and Academic Resources Coalition)

NUMERIC AND SPATIAL VISUALIZATION TOOLS

Data visualization, according to the Edinburgh Online Graphics Dictionary, is "The set of techniques used to turn a set of data into visual insight. It aims to give the data a meaningful representation by exploiting the powerful discerning capabilities of the human eye" (http://homepages.inf.ed.ac.uk/rbf/GRDICT/grdict.htm).

Although there are a range of commercial and academic domain-specific data visualization tools, there are also a number of collaborative web services using Web 2.0 technologies (including mashups or bricolage) that venture into the numeric and spatial data visualization arenas. The following data visualization tools can be regarded as open data utilities that function without the restrictions of their commercial or academic counterparts while retaining the ethos of other open initiatives, such as open source and open access.

- Data360 (www.data360.org) embraces the Web 2.0 concepts of participation and collaboration "to provide clear context on important cultural, environmental, social and economic issues."
- Many Eyes (http://services.alphaworks.ibm.com/manyeyes/home), an IBM utility, wants "to 'democratize' visualization and to enable a new social kind of data analysis."
- Swivel (www.swivel.com) aims "to liberate the world's data and make it useful so new insights can be discovered and shared."
- Gapminder (www.gapminder.org/downloads/applications), a Swedish foundation whose Trendanalyzer software was recently acquired by Google, contains only 16 variables. However, collaboration is planned with the United Nations Statistic Division to visualize millennium development goals with several World Development Charts in addition to visualizing the U.N. common database.

With the exception of Gapminder, registration allows users to upload their own data to these services with the understanding that the data is made freely available to all. If users wish to impose restrictions, such as using data within private groups or collaborations, then a fee is charged.

MASHUPS

Content used in mashups is typically sourced from a third party via a public interface or Application Programming Interface (API). There are literally hundreds of spatial data mashups available that can be created with very basic programming skills. Content from Web 2.0 services, such as Flickr photographs, can be georeferenced, plotted, and visualized using a range of mapping services, such as MS Virtual Earth, Google Earth, Yahoo! Maps, and NASA's World Wind.

ProgrammableWeb (www.programmableweb.com/tag/mapping) lists approximately 1,400 spatial mashups that utilize a whole range of Web 2.0 services. However, GeoCommons (http://geocommons.com) formalizes a spatial approach to data visualization. This utility allows users to upload, download, and search for spatial data; create mashups by combining data sets; and create thematic maps.

RESEARCH EXAMPLES OF SPATIAL MASHUPS

Web 2.0 technologies and interactive mapping products have paved the way for research organizations to explore and expose their findings in new and innovative ways:

- SRON, the Netherlands Institute for Space Research, and the KNMI, the Royal Netherlands Meteorological Institute, produced several data products via their SCIAMACHY Google Earth network (www.sron.nl/index.php?option=com_content&task=view&id=1506&Itemid=588).
- The British Oceanographic Data Centre wrote a Keyhole Markup Language (KML) generator application to automatically provide a KML file with each data request in order to enhance their spatial information (www.bodc.ac.uk/about/news_and_events/google_earth.html).
- NASA's Goddard Earth Sciences Data and Information Services Center established a portal which provides access to NASA imagery downloadable as KML files for importation into Google Earth (http://daac.gsfc.nasa.gov/googleearth/index.shtml).
- The U.S. National Snow and Ice Data Center offer Google Earth files that allow users to overlay a range of data-based images such as iceberg tracks, glaciers, and GLIMS ASTER glacier footprints onto a virtual globe (http://nsidc.org/data/virtual_globes).
- The USGS Earthquake Hazards Program displays real-time earthquakes and plate boundaries in Google Earth (http://

earthquake.usgs.gov/research/data/google_earth.php).

- AntWeb (California Academy of Sciences) have developed tools to facilitate the use of ants in inventory and monitoring programs and to provide ant taxonomists with access to images of type specimens. Users of Google Earth can now plot all the ants known to AntWeb on a 3D interactive globe of satellite images (www.antweb.org).

GRID-ENABLED DATA

It could be hypothesized that such visualizations, new data products, and practices produced by large research organizations are one of many converging precursors to a whole new mode of meta-research. For example, the Joint Information Systems Committee (JISC)-funded National Data Centres, EDINA and Mimas, are currently investigating access to their geospatial data services via the National Grid Service (NGS) using open interoperability standards.

It is likely in the future that large research organizations such as those in the examples above are not only grid enabling their data but utilizing Web 2.0 tools and technologies to enhance resultant output and create inter- and intra-disciplinary collaborations. This will make services and new resources available to a completely new and potentially cross-disciplinary audience within an e-research framework.

NEW FORMS OF DATA PUBLICATION

A tangible example of e-research activity is the work being carried out by particle physicists at CERN in Geneva. They have built the largest particle accelerator in the world, contained underground in a circular tunnel 27km in circumference, which will help them to delve into the nature of matter. This instrument speeds up particles to then smash them into other particles. The collisions generate vast amounts of data, which are gathered by highly sophisticated sensors to then be sent to dozens of data centers for analysis.

The data used and produced in e-research activities can be extremely complex, taking different forms depending on the discipline. In the hard sciences, such as biochemistry, data can take the form of images and numbers representing the structure of a protein. Data in Social Sciences could, for instance, contain an individual's attitudes toward politicians.

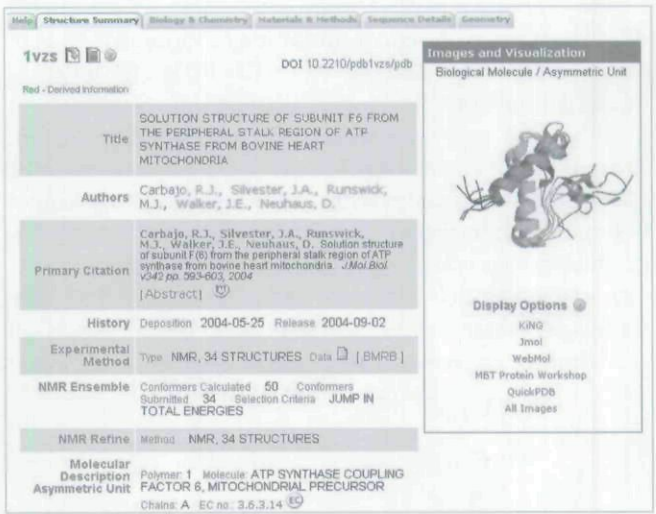
The e-infrastructure for this type of research activity, which includes the integration of data centers, collaborative environments, and grids, is currently being developed. Large resources are being committed to this—it has the potential of radically advancing knowledge with major implications for society at large.

E-RESEARCH AND LIBRARIES

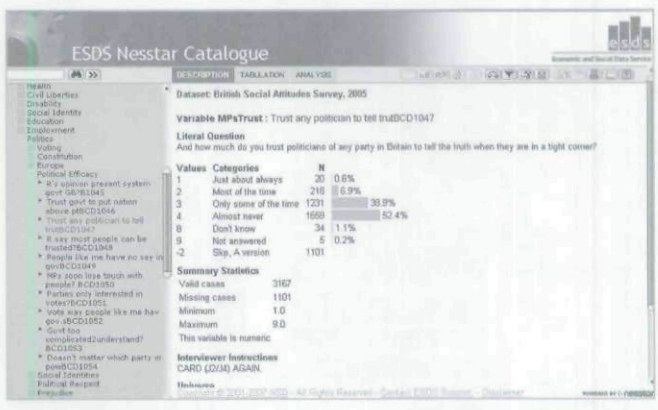
What is the relationship between e-research and libraries? Academic libraries have traditionally supported research by selecting, organizing, and making materials available for research purposes. However the role of libraries is changing and the road ahead remains unclear. Supporting e-research might be seen by some as a way forward for academic libraries, according to Anna Gold ("Cyberinfrasturcture, Data, and Libraries," *D-Lib Magazine*, September/October 2007; www.dlib.org/dlib/september07/gold/09gold-pt1.html and www.dlib.org/dlib/september07/gold/09gold-pt2.html).

Many groups are exploring how libraries can engage with e-research. In the U.K., the CURL/SCONUL Task Force on e-Research (www.nesc.ac.uk/esi/events/770/programme.cfm) has been examining librarians' understanding and awareness of e-research for the past few years. The Research Information Network (RIN) has surveyed researchers' use of academic libraries (www.rin.ac.uk/researchers-use-libraries) and consulted them about their view on the roles of libraries and e-research. Results from this survey revealed that there were significant differences of opinion between librarians and researchers on how library services should develop in the future. In the U.S., the National Science Foundation and the Association of Research Libraries have organized similar events to explore the new collaborative relationships.

A common outcome of the these discussions suggests that data curation is one role that libraries could take up in order to engage with e-research activities. Since the amount



Protein structure from RCSB Protein Data Bank (www.rcsb.org/pdb)



British Social Attitudes 2005 from ESDS (www.esds.ac.uk)

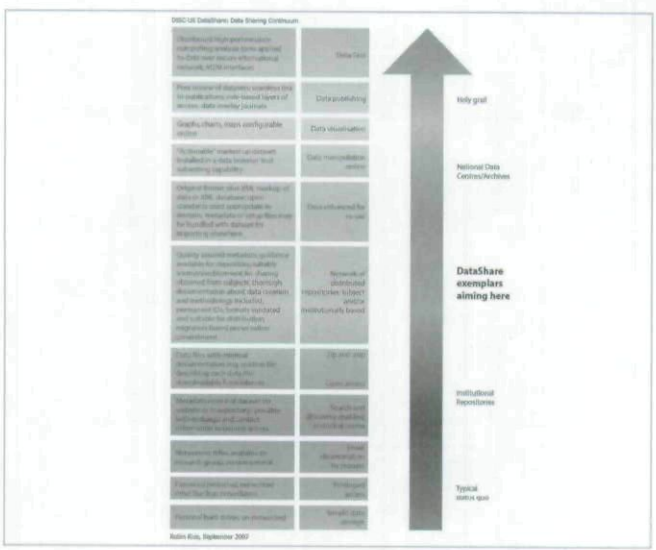
of data generated in the coming years will surpass all the data collected in human history, the management of this type of research output can only have major benefits and implications for future generations of researchers.

DISC-UK DATASHARE

In 2004, a group of data librarians and data managers formed the DISC-UK. This group has been working together to share data support experiences and expertise among four institutions: the Universities of Edinburgh, Oxford, and Southampton and the London School of Economics. Among other activities, some of its members started to explore the possibilities of curating data using repository technologies in 2004. This work resulted in a successful bid presented to JISC in autumn 2006 for a 2-year project, DISC-UK DataShare. This project aims to enhance the services at each institution by collaborating with IR managers while providing exemplars on a range of approaches and policies in which to embed the stewardship of data sets within DSpace, e-Prints, or Fedora-based digital repositories.

The Data Sharing Continuum attempts to enunciate the project's aims relative to current practice and people's expectations of what data sharing is about.

At the foot of the Data Sharing Continuum is the current typical scenario whereby many researchers are producing different types of data in a multitude of formats as part of their research. These are stored on CD-ROMs, flash drives, or on their personal computers with no information about how the data were captured or what they might represent. These data are not being curated; consequently they are at risk of being lost. Parallel to this, the open access movement has enabled libraries to establish institutional repositories in which to deposit, preserve, and provide open access to e-prints such as articles or theses produced at their own institutions. OpenDOAR (www.opendoar.org), a directory of academic open access repositories, lists 61 repositories worldwide containing data sets among their collection items. Nonetheless, most of those repositories have either no data set holdings or an insignificant number of data files with minimal documentation.



The data-sharing continuum

At the top of the continuum we find the national data archives such as the U.K. Data Archive, the British Atmospheric Data Centre or the Environmental Bioinformatics Data Centre (some of which, like the U.K. Data Archive, have been functioning for 40 years). These represent centers of excellence for the archiving and dissemination of data sets, the provision of online data manipulation tools with rich documentation and metadata that enhances the reuse of those datasets. This Holy Grail exemplifies the situation in which data used and produced during e-research activities is curated in a variety of interoperable systems, institutional repositories, and data archives, which allow distributed analysis over secure networks promoting and enabling reuse and repurposing of the data.

DataShare sits in the middle of the continuum, attempting to improve the current situation by making use of repository technologies to store data in standard formats with quality metadata, including information about the methodology used for their creation and also provides exemplars on how to anonymize data, obtain the consent from the creator to share his data openly and create data management plans, and commit to migration-based preservation of the data.

"Digital research data are products and by-products of the research process. They form an essential part of the evidence necessary to reconstruct and evaluate the results of research, and the events and processes leading to those results" (RIN, 2007).

This article has highlighted some of the key technologies, tools, and organizations within the open data movement. In an ideal world, governments, funding bodies, universities, nongovernmental and intergovernmental organizations, industry, and the general public would be encouraged to cohesively embrace such concepts—concepts that arguably have the potential to contribute to the social, cultural, and educational knowledgebase for the common good by not only increasing possibilities of knowledge transfer but ultimately in the generation of new knowledge itself.

During the coming months DISC-UK DataShare will participate in the current discussion in this area, attempting to bridge gaps between the communities and concepts while advocating for a world of open data. The future is bright. The future is open data!

[An abbreviated version of this article was presented at the Internet Librarian International Conference, London, Oct. 9, 2007, www.internet-librarian.com/programme.shtml. —Ed.]

Stuart MacDonald *(stuart.macdonald@ed.ac.uk) is data librarian at EDINA and Edinburgh University and* ***Luis Martinez Uribe*** *(l.martinez@lse.ac.uk) is digital repositories research coordinator, Oxford e-Research Centre. When this article was written, Uribe was data librarian at The London School of Economics and Political Science.*

Comments? Send email to the editor (marydee@xmission.com).