



New Shades of Grey: The Emergence of E-Science, Scientific Data and Challenges for Research Libraries

Julia Gelfand
University of California, Irvine
14 December 2009



Examples of charge to Library Data Team

Background:

- Data is an increasingly important resource for numerous academic programs on campus. Interest in data at a national and international level in academia is evidenced by the plans for a National Science Foundation cyberinfrastructure, and other e-science-related initiatives that are creating networked grids for computationally intensive research and collaborations (e.g. San Diego Supercomputer Center).

Charge:

- New Shades of Grey by Julia Gelfand, presented at the 11th International GrayLiterature Conference, Washington D.C., Dec. 14-15, 2009
- In addition to creating a Program Proposal, the Team will also develop a time line for the Program & a plan to implement the Program. The Team is also charged with monitoring the emerging issues and trends related to the provision of data services, data resources, and data standards--and identifying effective ways to share this information with library colleagues.

GL 2009 Themes

- The impact of Grey Literature on Net Citizens
- Uses and applications of subject based Grey Literature
- Grey Literature Repositories Revisited
- Open Access to Grey Resources



Simple Definition for eScience

- “The term ‘e-Science’ denotes the systematic development of research methods that exploit advanced computational thinking”
- “Such methods enable new research by giving researchers access to resources held on widely-dispersed computers as though they were on their own desktops. The resources can include data collections, very large-scale computing resources, scientific instruments and high performance visualisation.”



Some Examples of e-Science Projects

- Particle Physics
 - Global sharing of data and computation
- Astronomy
 - Virtual observatory for multi-wavelength astrophysics
- Chemistry
 - Remote control of equipment & electronic logbook
- Engineering
 - Nanoelectronics
- Healthcare
 - Sharing normalized mammograms, telemedicine
- Environment
 - Climate modeling



Interdisciplinary Challenges

- **Critical support for eScience**
 - **Clarifies distinctions in research methodologies**
 - **Not to be confused with multidisciplinary, transdisciplinary and other forms of disciplinarity-splicing**
 - **Supports Evidence-Based scholarship**
 - **Aligns the Clinical and Translational Sciences**
 - **Affirms new emerging directions and disciplines**



And from the UK:

- Is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it...and the purpose of the UK E-Science initiative is to allow scientists to do “faster, better, or different research.”

John Taylor, Director General of Research Council's Office of Science
& Technology



UK All Hands Agenda Dec 2009

- Social Sciences, Arts & Humanities
- Medical & Biological Sciences
- Physical & Engineering Sciences
- Environmental & Earth Sciences
- Sharing & Collaboration
- Distributed & High Performance Computing Technologies
- Data & Information Management
- User Engagement
- Foundations of eScience



Data Integrity Principle

- **Ensuring the integrity of research data is essential for advancing scientific, engineering and medical knowledge and for maintaining public trust in the research enterprise. Although other stakeholders in the research enterprise have important roles to play, researchers themselves are ultimately responsible for ensuring the integrity of research data**



Four recommendations come from this report

- 1) Researchers should design and manage their projects so as to ensure the integrity of research data, adhering to the professional standards that distinguish scientific, engineering and medical research both as a whole and as their particular fields of specialization.
- 2) Research institutions should ensure that every researcher receives appropriate training in the responsible conduct of research, including the proper management of research data in general and within the researcher's field of specialization. Some research sponsors provide support for this training and for the development of training programs.
- 3) The research enterprise and its stakeholders – research institutions, research sponsors, professional societies, journals and individual researchers – should develop and disseminate professional standards for ensuring the integrity of research data and for ensuring adherence to these standards. In areas where standards differ between fields, it is important that differences be clearly defined and explained. Specific guidelines for data management may require reexamination and updating as technologies and research practices evolve.
- 4) Research institutions, professional societies, and journals should ensure that the contributions of data professionals to research are appropriately recognized. In addition, research sponsors should acknowledge that financial support for data professionals is an appropriate component of research support in an increasing number of fields.



Definition of Grey Literature

“That which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers.”

GL Luxembourg Conference, 1997



e-Science Defined

“e-Science is not a new scientific discipline in its own right: ...is shorthand for the set of tools & technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways.” (Hey & Hey, 2006)

- **Cyberinfrastructure** – more prevalent usage of term in US
 - NFS: Revolutionizing science and engineering through Cyberinfrastructure, 2003 (Atkins Report)
 - Describes new research environments in which advanced computational, collaborative, data acquisition and management services are available to researchers through high-performance networks... more than just hardware and software, more than bigger computer boxes and wider network wires.
 - It is also a set of supporting services made available to researchers by their home institutions as well as through federations of institutions and national and international disciplinary programs
 - More inclusive of fields outside STM, emphasizes supercomputing & innovation



Cyberinfrastructure/ e-Infrastructure and the Grid

- The Grid is a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collection of individuals, institutions and resources' (Foster, Kesselman, and Tuecke)
- Includes not only computers but also data storage resources and specialized facilities
- Long term goal is to develop the middleware services that allow scientists to routinely build the infrastructure for their Virtual Organizations



Cyberinfrastructure Goals

- A Call for Action
- High Performance Computing
- Data, Data Analysis and Visualization
- Visual Organizations for Distributed Communities
- Learning and Workforce Development



Data Deluge?

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Google's founding philosophy is that we don't know why this age is better than that one: If the statistics...say it is, that is good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them...

Overview

- E-Science is the big picture
- Open Data is the goal
- Digital Repositories and Open Access are the methods
- Vision of 'joined up research' can be the process
- Combining cultures, connecting people; means new roles for libraries

e-Research

- E-Science is a shorthand for a set of technologies and middleware to support multidisciplinary and collaborative research
- E-Science program is ‘application driven’: the e-Science/Grid is defined by its application requirements
- There are now ‘e-Research’ projects in the Arts, Humanities, and Social Sciences that are exploiting these ‘e-Science’ technologies

What Really is e-Science?

Possesses several attributes – E-Science is:

- Digital data driven
- Distributed
- Collaborative
- Transdisciplinary
- Fuses pillars of science:
 - Experiment, Theory, Model/Simulation, Observation/Correlation

Chris Greer, 16.10.08



Context of E-Education

E-Education is:

- Information-Driven
- Accessible
- Distributed
- Interactive
- Context-Aware
- Experience and Discovery-Driven
- Potentially personalized



Beyond the Web: Open Source, Open Access

- Scientists developing collaboration technologies that go far beyond the capabilities of the web
 - To use remote computing resources
 - To integrate, federate and analyze information from many disparate, distributed, data resources
 - To access and control remote experimental equipment
- Capability to access, move, manipulate and mine data is the central requirement of these new collaborative science applications
 - Data held in file or database repositories
 - Data generated by accelerator or telescopes
 - Data gathered from mobile sensor networks



Orders of Magnitude

“In 2006, the amount of digital information created, captured, and replicated was 1.288×10^{18} bits (or 161 exabytes) ...this is about 3 million times the information in all the books ever written.”

Three years later, it far exceeds this.



Select Recommendations about Data Management

- Educate trainees and current investigators on responsible data sharing and reuse practices
- Encourage data sharing practices as part of publication and funding policies (NIH & other mandates)
- Fund the costs of data sharing and support for repositories



Key Drivers for e-Science

- Access to Large Scale Facilities and Data Repositories
 - e.g. **CERN LHC, ITER, EBI**
- Need for production quality, open source versions of open standard Grid middleware
 - e.g. **OMII, NMI, C-Omega**
- Imminent ‘Data Deluge’ with scientists drowning in data
 - e.g. **Particle Physics, Astronomy, Bioinformatics**
- Open Access movement
 - e.g. **Research publications and data**



Key Elements of a National e-Infrastructure

1. Competitive Research Network
2. International Authentication and Authorization Infrastructure
3. Open Standard Middleware Engineering and Software Repository
4. Digital Curation Center
5. Access to International Data Sets and Publications
6. Portals and Discovery Services
7. Remote Access to Large-Scale Facilities, e.g. LHC, Diamond, ITER,...
8. International Grid Computing Services
9. Interoperable International Standards
10. Support for International Standards
11. Tools and Services to support collaboration
12. Focus for industrial Collaboration



Digital Curation Centers (DCC)

- Identify actions needed to maintain and utilize digital data and research results over entire life-cycle
 - For current and future generation of users
- Digital preservation
 - Line-run technological/legal accessibility and usability
- Data curation in science
 - Maintenance of body of trusted data to represent current state of knowledge in area of research
- Research in tools and technologies
 - Integration, annotation, provenance, metadata, security...



Digital Preservation: Issues

- Long-term preservation
 - Preserving the bits for a long time (“digital objects”)
 - Preserving the interpretation (emulation/migration)
- Political/social
 - Appraisal – ‘What to keep?’
 - Responsibility – ‘Who should keep it?’
 - Legal – ‘Can you keep it?’
- Size
 - Storage of/access to Petabytes of data
- Finding and extracting metadata
 - Descriptions of digital objects


Data Publishing: Background

- In some areas – most notably biology – databases are replacing (paper) publications as a medium of communication – think Genome mapping
 - These databases are built and maintained with great deal of human effort
 - They often do not contain source experimental data – sometimes just annotation/metadata
 - They borrow extensively from, and refer to, other databases
 - You are now judged by your databases as well as your (paper) publications
 - Upwards of 1000 (public databases) in genetics

Data Publishing: Issues

- Data integration
 - Compiling data from various sources
- Annotation
 - Adding comments/observations to existing data
 - Becoming a new form of communication
- Provenance
 - ‘Where did this statistic come from?’
- Exporting/publishing in agreed formats
 - To other programs as well as people
- Security
 - Specifying/enforcing read/write access to *parts* of your data

Considerations



Massive Data Sets
means
Federation,
Integration,
Collaboration

There will be more
scientific data
generated in the next
five years than in the
history of humankind

Evolution of many-core
& multicore activities
Parallelism
everywhere

What will you do with
100 times more
computing power?

The power of the
client + cloud =
Access Anywhere,
Anytime

Distributed, loosely-
coupled, applications
at scale across all
devices will be the
norm



The 'Cosmic Genome' Packages: Examples

- World Wide Telescope (<http://www.worldwidetelescope.org>)
- The Sloan Digital Sky Survey (www.sdss.org)

Valuable New Tools

- GenePattern (<http://www.codeplex.org>)
- GalaxyZoo (<http://www.galaxyzoo.org>)
- Semantic Annotations in Word
- Chemistry Drawing for Office
- New Models of Scientific Publishing
 - **Have to publish the data before astronomers publish their analysis**
 - **Integrates data and images into research papers**

Emergence of a Fourth Research Paradigm

1. Thousand years ago – Experimental Science
 - **Description of natural phenomena**
2. Last few centuries – Theoretical Science
 - **Newton's Laws, Maxwell's Equations...**
3. Last few decades – Computational Science
 - **Simulations of complex phenomena**
4. Today – Data-Intensive Science
 - **Scientists overwhelmed with data sets for many different sources**
 - Data captured by instruments
 - Data generated by simulations
 - Data generated by sensor networks
 - **e-Science 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



Cloud Science Examples

- Using different tools & software to demonstrate the value of cloud services
- Scientific Applications on Microsoft Azure
 - Virtual Research Environments
 - Oceanography Work Bench
 - Project Junior @ Newcastle University

The background features a collage of digital and office-related icons. On the left, there's a keyboard with keys like 'P', 'Q', 'R', 'U', 'V', and 'W' visible. Below the keyboard, there's a 'data' icon (a folder with a document) and a 'save' icon (a floppy disk with an arrow). The overall color scheme is warm, with shades of orange and yellow.

Collaborative Online Services

- Exchange, Sharepoint, Live Meeting, Dynamics CRM, Google Docs, etc.
- No need to build your own infrastructure of maintain, manage servers
- Moving forward, even science-related service could move to the Cloud (e.g. RIC with British Library)

A world where all data is linked...

- Data/information is inter-connected through machine-interpretable information (e.g. paper X is about star Y)
- Social networks are special case of 'data meshes'
- Important/key considerations
 - Formats or "Well-known" representations of data/information
 - Pervasive access protocols are key (e.g. http)
 - Data/information is uniquely identified (e.g. URLs)
 - Links/associations between data/information

How is it done?

- e-Science

- Science increasingly done through distributed global collaborations enabled by the Internet
- Using very large data collections, terascale computing resources & high performance tools

- Grid

- New generation of information utility
- Middleware, software & hardware to access, process, communicate & store huge quantities of data
- Infrastructure enabler for e-Science

- Cloud

- New, easier & cheaper opportunities to host, store, share & integrate, tag & link; utilizes more Web 2.0 applications



More Implications of Technology Trends

- Web 2.0
 - More egalitarian – affects scientists, students, educators, general public
 - Collaborative classification – flickr
 - Power of collective intelligence – Amazon
 - Alternative trust models – Open Source
- Service Orientation
 - within & outside of libraries
- Semantic Web
 - Promotes linking



Data Centric 2020

Data-centric 2020 vision resulting from Microsoft 'Towards 2020 Science' (2006)

Data gold-mine

'Multidisciplinary databases also provide a rich environment for performing science, that is a scientist may collect new data, combine them with data from other archives, and ultimately deposit the summary data back into a common archive. Many scientists no longer 'do' experiments the old-fashioned way. Instead they 'mine' available databases, looking for new patterns and discoveries, without ever picking up a pipette.'

'For the analysis to be repeatable in 20 years' time requires archiving both data and tools.'



Organizing eScience Content: Examples

Tags

- Subject
- Instrumental
- National
- National/Subject
- International
- Regional
- Consortia
- Funding Agency
- Project
- Conference
- Personal
- Media Type
- Publisher
- Data Repositories

Project Meaning or Origin

- arXiv, Cogprints
- University Research Institutes – Southampton, Glasgow, Nottingham (SERPA). Max Planck
- DARE (all universities in the Netherlands), Scotland IRIS)
- OceanDocsAfrica
- Internet Archives 'Universal,' Oaister (Harvester)
- White Rose UK
- SHERPA-LEAP (London E-Prints Access Project)
- NIH (PubMed), Wellcome Trust (UK PubMed), NERC (NORA)
- Public Knowledge Project Eprint Archive
- 11th Joint Symposium on Neural Computation, May 2004
- Peer to peer
- VCILt Learning Objects Repository, NTSDL (Theses), Museum Objects, Repositories, Exhibitions
- Journal Archives
- UK Data Archive; World Data Centre System; National Oceanographic Data Center (USA)

Ongoing Issues for e-Science

- Macro and micro issues are similar for both text and data repositories
- IP and Licenses**
- Distributed over many researchers
- Over national boundaries
- Lack of awareness amongst researchers
- Cultural roots and resistance to change
- Funding costs, sources & accountability
- Politics – institutionally & within the disciplines
- Standards
- Interoperability
- Vocabularies & Ontologies

Research Issues:

- Information retrieval
- Information modeling,
- Systems interoperability, and policy issues associated with providing transparent access to complex data sets

**Necessary to understand science practices: technical, social & communicative structure in order to adapt licensing solutions to the practice of e-science



New Roles: Data Scientist

New Skills Required

- Understanding of basic research problem & interdisciplinary connections
- Quantitative & systems analysis
- Data Curation & Text Mining
- Integrate data management within the LIS curriculum
- Stronger IT & negotiation skills
- Deeper subject backgrounds; standards & resources

In Practice

- Various approaches to develop and obtain digital curation skills
- Established ties to faculty
- Skills are there but often in discrete communities: we need to bring communities together
- Integration within the curriculum: undergraduate students library & information science, archival studies. Computer science
- Provide recognition and career path for emerging 'data managers & scientists'

There must be a blurring of the boundaries between previously well defined silos that existed between information managers and data managers



Role for Libraries in Digital Data Universe

- Data as primary source material – Libraries
 - Will not be primary providers of large scale storage infrastructure required
 - Will not provide the specialized tools to work with data
 - Will not provide the detailed information about the data
 - Unlikely to provide the solutions to digital preservation because of cost
- Can contribute library practices
 - Collection policies (appraisal, selection, weeding, destruction, etc.)
 - Data clean up, normalization, description
 - Data citation
 - Curation and preservation
 - Collaboration with researcher re scholarly communication, deposit, education, and training
 - Innovative discovery and presentation mechanisms
- Data part of ‘enhances publication’ – Libraries:
 - **Well positioned to define standards for**
 - Taxonomies and ontologies (for complex publications that include data)
 - Persistent identifiers
 - Consistent description practices
 - Data structuring conventions
 - Interoperability protocols for searching and retrieval
 - **Well positioned to exploit IR experiences**



Role of Digital Libraries - IRs

- Institutional Repository is a key component of e-Infrastructure
 - Mostly in library domain
 - Access and preservation
 - Digitization – data archaeology
 - Interoperable with departmental, national, subject repositories
- Data Curation
 - Creation metadata, preservation institutional intellectual assets
 - But disparate data types and ontologies
- Training Provision
 - Research methods training for researchers
 - Data creation, documentation, managements
- Advocacy, policy setting
 - Cross disciplinary approach to key issues
 - Expand OA agenda
 - Interweave e-Research, OA, and
 - Virtual Research Environments

Roles for Libraries

- Institutional Repositories accept “small” datasets (size of subject outside remit of Data Repositories). Data deposited in IR until accepted by data repository
- Development of Regional or Discipline Repositories alongside IRs (singly or via consortia). Research libraries a natural home for content curation, (with funding)
- Mapping of commonalities (e.g. metadata) across disciplines, maintaining ready interoperability
- Management of metadata throughout a research project
- Address conditional and role-based access requirements for scientific data
- Support e-Science interface functions for local users
- Adding Value: linking, annotation, visualization
- Libraries and researcher can add value by creating ‘e-Science Mashups’ - data needs to be re-used in multiple ways, on multiple occasions and at multiple location (reuse, remix)

Reinventing the Library

Intensity urgently needed to support eScience:

- Emphasis is Data – thus, new forms of collections & auxiliary resources
- Institutional commitment
- Sustainable funding models
- Redefining the library user community-include research
- Legal and policy frameworks
- Library workforce skills – infusion for data science management
- Library as a computational center as well as a text & media center
- Sustainable technology framework



Academic Research Libraries

“It is the research library community that others will look to for the preservation of digital assets, as they have looked to us in the past for reliable, long-term access to the ‘traditional’ resources and products of research and scholarship.”

Association of Research Libraries (ARL) Strategic
Plan 2005-2009



Thank You

Questions / Comments?
jgelfand@uci.edu