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

By

Julia Gelfand
Applied Sciences & Engineering Librarian
University of California, Irvine

Presentation at the 11th International Conference on Grey Literature
Library of Congress
Washington, DC

14 December 2009

#1 Abstract: Four of the themes of the Eleventh International Conference reinforce the subject of this paper. They include 1) the impact of GreyLiterature on Net Citizens, 2) uses and applications of subject based Grey Literature; 3) Grey Literature Repositories Revisited; and 4) Open Access to Grey Resources. E-Science is among the latest forms of literature that has largely been born in the public domain and funded with federal dollars to explore large scale computing potential for the challenges of basic research in the physical sciences, life sciences, medical and clinical sciences, engineering, information sciences and technology, and most recently as it influences public policy. This emergence of eScience demonstrates that the data poses new challenges in its needs to find a home that is safe, open for repurposing, reuse and additional applications. Data can be classified in many ways utilizing a range of appropriate metadata descriptors to enhance its utility for different and future applications. Statisticians and scientists have been concerned about how this data can be retained, archived, preserved and entered into collections that honor its safekeeping and potential for future manipulation. Unlike the information products that are derived from its applications, data is often raw, formulaic, rough, distributed, numeric, tabular and/or loose. Computer scientists and technologists have explored grid computing and now cloud computing to realize its utility as new components in E-Science. This new frontier is among the latest efforts to fuse transdisciplinary, collaborative, distributed pillars of science into a mix of experiments, theories, models, simulations, observations and correlations, which I will demonstrate are new forms of grey literature. The full lifespan of eScience demands that it configures the data curation and preservation aspects and extends the lifetime of grey literature to new challenges, which print and text never experienced. How libraries will cooperate with scientific communities in realizing this new potential and take responsibility for this aspect of grey literature is most curious in this era that observes a fast track maturation of eRepositories around the globe, many external competing demands while trying to effortlessly anticipate and respond to challenges in scholarly communication, open access and the public's right to know. This paper addresses these and related issues of eScience and science librarianship, within the realm of grey literature at a time of institutional and scientific competitiveness and economic uncertainty.

My interest in this new form of grey literature evolved in parallel to when I became aware of eScience and began to see the themes of this conference as fusing together instead of remaining separate silos. The time period was approximately late 2001 when the first major meetings were taking place on eScience, attended by librarians and information scientists, first in the UK and then in the US. In this short period of less than five years a litany of meetings have been held and the role has firmly been on how to collect, retain and archive the different types of data that accompany large sponsored eScience research projects. I have attended several meetings in recent years and observed how the role of eScience has emerged and join colleagues around the globe in learning how academic research libraries are responding to the call for participating in some new eScience initiatives which has proven both challenging and problematic. This paper advances some thinking that followed a paper that I

presented in Beijing in February 2009 on the "The Changing Collections for eScience: What it Means for Libraries."[1] I concluded that eScience is not the anticipated extension of digital library development, but requires examining how to handle large data sets from disciplines in which liaison librarians may not be sufficiently expert or prepared to lead library programs in this field. Increasingly, a new venture with scientists and IT colleagues has been launched, libraries are indeed acquiring some of the background necessary to establish appropriate ties to institutional programs in which cyberinfrastructure has an important role. By making experimental data more shareable and applying the same tenets of organizational structure, entire areas of new cooperation and collaboration emerge. This may include metadata, storage, archiving and preservation, and the needed public service elements for repurposing.

#2 Many research libraries have formed task forces, teams and committees to study how to manage this data and work with campus leadership to ascertain how to support the data from these mammoth projects.

#3 My paper explores not just the new roles libraries are adopting to further eScience scholarship and how library collections will change, but addresses four interrelated themes of this conference:

1) the impact of GreyLiterature on Net Citizens,

2) uses and applications of subject based Grey Literature;

3) Grey Literature Repositories Revisited; and

4) Open Access to Grey Resources. Each of these themes, I propose is an undercurrent in the work and thinking of most science librarians today.

#4 To be on a common page, the simple definition of eScience that I choose to use is: "The term "e-Science" denotes the systematic development of research methods that exploit advanced computational thinking."[2] It is further amplified by this clause, AND "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."

#5 The landscape for eScience and Large Data Sets is quickly changing. Today, there is much work in this field and many examples to cite. There are endless examples of eScience collaboration and product development. As indicated they span all disciplines and multiple intersections between them. The new byproducts increasingly have commercial value, intellectual property that must be protected and technology transfer potential.

#6 Interdisciplinary challenges confront eScience at every turn. With competing interests from the physical sciences as well as increasingly from the life and biomedical sciences, engineering and technology, developing an open mind to different approaches is critical. Interests in the social, economic, political and policy implications complicate issues further. A summary of this can be seen on this slide but some viewpoints illustrate the variations of nuance:

> "…the most formidable challenges in the revolution of biomedical research…require expertise from disciplines that fall at the boundaries of most AMCs [medical schools] and their parent universities. These include the disciplines that encompass the applied & theoretical physical sciences, including engineering, social sciences, law & business. As intellectual campfires of interdisciplinary research & learning form, significant barriers born from entrenched institutional practices and culture surface, and if unchecked, these barriers may drown the flames of discovery."[3]

> "Sharing biomedical research and health care data is important and difficult …Many initiatives fund, request or require researchers to share their data. These initiatives address the technical aspects of data sharing but rarely focus on incentives for key stakeholders."[4]

My apologies if I omit something with which you are already familiar or think of as very mainstream in this dynamic environment.

Several factors lead one to realize why eScience is critical to the contemporary academic research library landscape. "It has been speculated that the Internet revolution is less than 15% complete, based on the number of users, total bandwidth, total amount of content, number of devices, number of applications.[5] The history of eScience is relatively young, with its early days in 1966 when the US Government's ARPANET was developed to explore methods for "resource sharing among computers." Since then, a quiet coexistence takes place with continued government exploration from several agencies and departments.

eScience has had relationships with eSocial Science since the early 1980s with the focus on quantified methods and how to handle those large data sets, commonly associated with ethnographic research, public opinion polling, census data, voting behavior and obviously other examples of grey or fugitive literature. One way that much of this has been treated is best captured by the work of the ICPSR (Interuniversity Consortium for Political and Social Research, University of Michigan)[6], which provides access to a vast archive of computable readable data in social science research and the tools to make it compatible with statistical software so the data can be manipulated as researchers prefer.

Formal and informal discussions about eScience began to take place in the US around 1999 with the early meetings devoted to it since 2003 when organizations composed of the research library community such as the Coalition for Networked Information, the National Science Foundation, Association of Research Libraries and other consortial organizations began to work with the scientific communities and national research laboratories.

#7 E-Science activities were also unveiling themselves in Europe, particularly in the UK under the Director General of the Research Council's Office of Science and Technology, and under whose auspices many of the first early discussions took place such as the first eScience All Hands Meeting in Nottingham in September 2005. Within a very short interval, the International Workshop creating the Information Commons for eScience was hosted by Science Commons at UNESCO Headquarters in Paris also in September 2005, and the Eighth International Bielefeld Conference in Germany took place in February 2006 with sessions devoted to eScience and subsequently many other conferences have taken place.

#8 Just last week, the fifth annual UK All Hands Meeting took place in Oxford in conjunction with the IEEE e-Science meeting and the UK Research Council's Review on eScience, giving the opportunity to bring the UK community together with the international leaders in eScience and its assessment and funding bodies, and we look forward to the proceedings of that meeting in coming days.

The global activity in eScience is accelerating but my focus is indeed American-centric and I excuse that as a shortcoming in this international context, but that is the work environment I know best. The role of the digital library movement has certainly motivated me to see new potential, but I caution you to understand that eScience is not entirely synonymous with creating digital resources and collections, work of which academic libraries today are fully engaged.

The programmatic lineup of this meeting in Oxford illustrates the fast maturity of eScience as it reflects the following themes, again confirming the importance of these recommendations to ensure data integrity and preservation:

- 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

A very recent report, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*[7], summarizes the data management challenges facing the scientific research community. It discusses implications for big and small science, and was issued by the Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, of the National Academy of Sciences, within the last few months. In its Executive Summary, it states that, "The integrity of data in a time of revolutionary changes in research practice is too important to be taken for granted. Consequently, this report affirms the following general principle for ensuring the integrity of research data:

#9 **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.**[8]

#10 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.[9]

The strands of grey are very evident in these recommendations and as my paper emphasizes the data collection, curation, archiving, enhancement and control, reuse and functionality that reaffirms the collaborative approach to eScience. I cannot help but be reminded of the cautions about the excesses and largesse of big business in recent years as we encourage business to be carried on in a moral manner while insisting that a free and open market benefit democracy. These perceived "old-fashioned virtues" may apply to how eScience parallels such lofty achievements by giving new content to old ideals, where the "only abiding thing is change."[10] This report will also serve as the framework and programmatic theme of the forthcoming International Association of Technological Libraries (IATUL) meeting in June at Purdue.

#11 Definition of Grey Literature) Still today, I focus on the cursory landscape of just a few of these themes as they reflect the tone, syntax and mixed flavors of greyness as they are known to us. For context let's utilize the current and now standard definition of grey literature adopted at the 1997 Luxembourg conference, "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."[11] Does data not fit this bill?

#12 Many programs and experiences put and retain eScience on my radar, which includes elements of all these concepts in this wordle scattergram. My close alignment with the Grey Literature movement for nearly two decades which has certainly been at the forefront of digital publishing, handling and treating statistics and data, has born witness to the changed roles due to online collaboration, sharing and distribution. The hues of grey have greatly blurred as grid computing began to capture the way science could be collected, saved, and repurposed. The early work at CERN and by astrophysicists and how they adopted the applied physics preprint server, ArXiv (http://arxiv.org/) begun in 1992 by Paul Ginsparg, formerly of Los Alamos and now at Cornell University.

#13 It was from conferences at CERN and supporting research at my institution that I became familiar with the vocabulary of terms like cyberinfrastructure (insert old SLIDE #4,) computational applications

that a more accurate definition of eScience emerges, and more recently the potential of grid and cloud computing, often introduced by our colleagues at these Grey Literature meetings with the early and ongoing work by Keith Jeffrey and Anne Asserson and last year's paper by Elly Dijk and colleagues on "Accessing Grey Literature in an Integrated Environment of Scientific Research Information" and the work of the Dutch DARE programme[12] and figuring out how massive projects like the human genome databank pioneered and drove our current thinking.

#14 Understanding what cyberinfrastructure means is essential to grasp the potential for eScience. "At the head of the cyberinfrastructure vision is the development of a cultural community that supports peer-to-peer collaboration and new modes of education based upon broad and open access to leadership computing: data and information resources; online instruments and observatories; and visualization and collaboration services. Cyberinfrastructure enables distributed knowledge communities that collaborate and communicate across disciplines, distances, and cultures. These research and education communities extend beyond traditional brick and mortar facilities, becoming virtual organizations that transcend geographic and institutional boundaries.

#15 This vision is new, exciting and bold.[13] Five elements direct the NSF vision:

1. A Call for Action
2. High Performance Computing
3. Data, Data Analysis and Visualization
4. Visual Organizations for Distributed Communities
5. Learning and Workforce Development[14]

As information technologies and computing power become ever more powerful and easy to use, the new frontier of science and engineering is to harness the collective power of our researchers on a global scale. Some disciplines are far along in the process of sharing information and creating fully collaborative infrastructure to power innovation; others are on the verge of embracing the power of cyberinfrastructure to transform their research.

Building on the short history, of eScience one has to understand the thematic intersections established by these events that are important to both libraries and to science:

- Coalition of Networked Information, (CNI) – from 2002 significant parts of their meetings have advanced the concepts of eScience and data within the realm of libraries and data centers
- CNI, together with the Association of Research Libraries (ARL) co-sponsored a joint sponsored conference in Washington, DC in October 2008 on "Reinventing Science Librarianship."[15] The emphasis of that meeting was to examine how the research library community can better participate in the research enterprise on the campuses where eScience is taking place. One of the keynote addresses at that conference was delivered by Chris Greer, Director of the National Coordination Office, Networking and Information Technology Research and Development Program (NITRD) who opened his remarks with this passage from Chris Anderson that was quoted in *Wired Magazine*:

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

> Google's founding philosophy is that we don't know why this page is better than that one: if the statistics …say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data Google can translate Klingon into Farsi as easily as it can translate French into German).[16]

- American Association for the Advancement of Sciences (AAAS) also influenced me about collaborative and participative roles that libraries can have with scientists. This was reinforced at the 2009 Annual Conference where I began to understand the latest developments in cloud computing and how that potential will offer eScience enormous opportunities and saw direct applications to the work in genomics, astrophysics, and other applications. It was there that I had the chance to engage in discussions with contributors to eScience such as Dr. Tony Hey, Vice-President of Technical Computing at Microsoft, Dr. Christine Borgman from UCLA and others.

#17 – (overview) It was Hey who wrote in 2005, "In the future, frontier research in many fields will increasingly require the collaboration of globally distributed groups of researchers needing access to distributed computing, data resources and support for remote access to expensive, multi-national specialized facilities such as telescopes and accelerators or specialist data archives."[17] He goes on to say, "There is also a general belief that an important road to innovation will be provided by multidisciplinary and collaborative research… ."

#18 - Hey predicted in 2005 that "Robust middleware services will be widely deployed on top of the academic research networks to constitute the necessary 'e-infrastructure' or 'cyberinfrastructure' – to provide a collaborative research environment for the global academic community. Hey continued, "This technology is likely also to change the nature of scientific publication with institutional or subject repositories linked to digital archives containing the primary research data."[18] For libraries, that was the gospel of just four years ago.

#19 – So, what actually is eScience and where is it today? Greer asked us at the meeting just over a year ago to imagine a world where…all of the text in all of the libraries worldwide is in a storage device – in your pocket; and the network responds at the speed of light to a plain language question with a perceptive answer and ….your contact lenses merge your digital and physical worlds."[19] Simply stated and confirmed, eScience is:

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

#20 Within the context of eLearning and eEducation, what has emerged on the horizons of distance education, education and research is more about communication and collaboration and the flattening of the world. We see the following attributes, also commonly associated with grey literature.

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

#21 The theme of Open Access for data has similar roots as it does for science information in general and grey literature in particular. The digital divide applied to scientific information has been reduced by different programs with publishers and public agencies that have directed access to commercial and society publications in developing nations and the emphasis we have on global and public health, vaccination, clean drinking water and eradicating diseases with the aid and support of foundations and international agencies. None of this could take place without understanding large data sets, mapping strategies, and reducing barriers to information. Dr. Harold Varmus, Nobel Laureate, former Director of the National Institutes of Health under President Clinton, proponent of PubMedCentral, one of the founders of the Public Library of Science known as PLoS, an evangelist for open access, and now President of the Memorial Sloan-Kettering Cancer Center in New York, reflects, "The Internet and the desktop computer have transformed the way science is practiced in virtually all fields, and the biomedical sciences have not been exempt. Today most scientists obtain, use and produce information in ways that would have been unrecognizable 20 or 30 years ago. These changes have profoundly and most contentiously affected an aspect of science that is at the core of any scientist's life: the ways that we publish, disseminate, read, store and retrieve research papers."[20] He goes on to write, "Viewed together, public digital libraries and open access publishing promise great benefits for science and society: equity, through universal and unfettered delivery of knowledge, mostly a product of public funding; more effective practice of science; and reduction of overall costs."[21] He does not allude to eScience specifically beyond proselytizing best practices for scholarly communication, but one can assume that his support for eScience is unequivocal as he notes the progress being made in making science and biomedical information available to the public at large in developed countries of the world and less-developed nations and the need for public digital libraries.[22] Just being aware of and sensitive to what the challenges are for science in the developing world is joining this optimistic horizon.[23] Open access has made enormous progress but still is not ubiquitous in the science arena. The PubMedCentral concept of depositing content in central, disciplinary, and institutional repositories allows for maximum flexibility and crawling of search engines and delivery of content at time of need. It is not perfect but reduces the burden of access.

#22 Scaling the science is important to understand and what this means for computing…"In 2006, the amount of digital information created, captured, and replicated was 1.288 x 10 to the 18^{th} bits (or 161 exabytes)…this is about 3 million times the information in all the books ever written." And, today…it can be estimated to add about another 20%

#23 Making a difference in working with data may include forming new partnerships both within institutions and cross-institutions and forming new consortia to achieve new goals for data management. Scaling and sharing costs, creating talent pools, tricks of the trade and training for the next generation are all needed. Selective recommendations for changing the culture and practices about managing data may include how to:

- Educate trainees and current investigators on responsible data sharing and reuse practices
- Encourage data sharing practices as part of publication and funding policies
- Fund the costs of data sharing and support for repositories[24]

"The National Center for Research Resources will continue to support and address integration of informatics research and solutions in all its programs and centers…NCRR will pursue integration across various domains of knowledge and research within NCRR and its partners at other NIH Institutes, other federal agencies, industry and foundations. Issues of integrity, durability, availability and security of data will continue to lay a critical role in this era of fast moving technologies and analysis tools." [25]

#24 There are some key drivers for eScience which have specific sources, expertise, applications and outputs – Such examples include:

- 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
  - **To research publications and data**

#25 A careful and long list of necessary concerns and needed components of the required national e-infrastructure are demonstrated. These are critical to the success of any data management and data sharing experience.

#26 Central to this is an understanding of Data Curation issues and challenges. Among the experts in this is Sayeed Choudhury, at the Johns Hopkins University Library and he collapses this into key considerations:

- Work with existing scientific systems
- Consider gateways for these systems as part of infrastructure development
- Focus on both human & technical components of infrastructure
- Human interoperability is more difficult than technical interoperability
- Trust[26]

The work at Johns Hopkins is definitely something to follow as is the Distributed Data Curation Center (D2C2) at Purdue.

#27 – Related to Data Curation is Data Preservation, perhaps closer on the comfort zone for libraries to work on as they have been committed to building digital depositories of journals, now books and other grey objects for a significant period of time.

#28 – The compilation of the preserved data suggests a new stream of information products and can best be described as "data publishing," or the creation of specialized databases. This may be a new form of open access products. eScience is probably too complex for the average citizen but to the academic or scientist familiar with the scale and scope, properties of the data, these resources will be of potential value. An example of when or why may be the scenario when a scientist is at a small lab and only needs a fraction of the data that was collected or is not part of an infrastructure that can support such a study then they can participate in reviewing or applying the data as needed when in the past they would be unable to have access.

#29 – They may include the issues of :

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

#30 Organizing these tasks in order to create appropriate metadata and future use of the data to either replicate, review, repurpose, refine or evaluate science, requires these considerations and contributes to the expanding universe of grey literature.

#31 Examples of these interdisciplinary and large scale computing efforts include "Cosmic Genome Projects" such as these noted

- World Wide Telescope (www.worldwide telescope.org) – Social Media Virtual Sky for science & education; Harvard/JHU/Microsoft collaboration with Goodman & Szalay
  - As of late Jan 09: 1,606,950 unique users
  - Average # of new users daily = 3,773
- The Sloan Digital Sky Survey (www.sdss.org) is the first major astronomical survey project – Gray & Szalay build public 'skyserver' archive for survey at JHU:
- **5 color images of ¼ of the sky**
- **Pictures of 300 million celestial objects**
- **Distances to the closest 1 million galaxies**

For a broader audience or a more average net citizen, we are seeing many new tools that allow researchers to work with the data and encourage broader utilities by students even at the K-12 levels and the public at large which have all extended the limits on what is possible and forced libraries to understand the collaboration inherent in this work and

- GenePattern, work at the Broad Institute at MIT (www.codedata.org) ,
- Galaxy Zoo (www.galaxyzoo.org)

#32 Tony Hey writes about the emergence of the Fourth Paradigm to accommodate eScience. You can see how data has dictated new challenges and outcomes.[27]

#33 Along with other proponents like, Michael Nelson, Hey introduces how cloud computing is a potential solution for requiring less supercomputer resources and more desktop and other developments now associated with Web 2.0 technologies. "Academic grids are a prototype of the cloud which also manifests as many forms of social media and can be more powerful than supercomputers and is predicted to be as influential as E-Business." [28]

#34 The emphasis on collaboration, sharing and communication which reinforces the Fourth Paradigm and demonstrates the utility of these powerful tools that we want to remind ourselves can run on the most basic laptop.

#35 We enter a universe where all data is linked and it should be clear to us as librarians why this is so important.

#36 We are reminded that it is still basic eScience, with the grid and now the Cloud that allows us to realize the potential we have for making eScience so vital today.

#37 With Web 2.0 services so critical to our users Cloud Computing is a likely alternative.

#38 The Data Center Gold Mine is yet another reminder how critical digital preservation and archiving is. These curatorial functions make it possible for data sharing to take place.

#39 The mature players in eScience have already come up with tags and organizational control to serve as the metadata criteria for identifying different parts of the content.

#40 Libraries are examining in a concerted way how to handle and treat eScience and be the objective and good steward that is essential for success in these kinds of eScience enterprises. The list of ongoing issues has been echoed by nearly everyone who has been associated with any of these projects. The one thing to discern is that these are all "works in progress" and are likely to go through many changes before standards are fully in place. Also, distinguishing between digital library initiatives, ePublishing issues, services and products is helpful because some of the players may be the same, but usually they are different, they need different skill sets and have very different priorities.

#41 Two of the most challenging situations is that librarians may not have the confidence to either participate at the level necessary or to realize that the library should be involved in what has been called "the eScience revolution" that will put libraries and repositories center stage in the development of the next generation research infrastructure."[29] Library School curriculum must begin to offer courses in data and it is understood that at UCLA in the Masters degree in Library and Information Science that will soon happen, however I am not sure where else that is on the horizon. The University of North Carolina is reporting more efforts to prepare graduates for data oriented roles as data consultants, distributors and managers and Catherine Blake at UNC says "that not everyone will have every skill… that librarians must forge deeper relationships with faculty and think like someone within a discipline…have strong negotiation skills and a knowledge of standards and resources." She concludes, "The roles exist, its not

clear where they will live within an institution."[30] There are several intensive institutes being offered by the University of Illinois[31] and probably elsewhere. The "Embedded Librarian" model may be a top slicing way to describe the effort but it is clearly needed. Many institutions have appointed specialist librarians with titles of Bioinformatics Librarian in the Health Sciences or Science Library where they are closely associated with the campus research enterprise.

#42 (may be most important slide) **Some** of the new roles specifically include these tasks, not commonly found in all libraries. But if we want to be involved in this new partnership, this is part of what we will have to do. Librarians are expected to continue to perform roles in anticipating and evaluating information needs, identifying appropriate sources and instructing about their uses and utility and making sure that they are retained and preserved as needed, but those traditional roles are indeed very different than handling data. The major difference is that between bibliographic and data-centric collections.

#43 **In conclusion**, there are many new roles for libraries as they develop repositories and build on their work in scholarly communication and open access.

#44 With the focus on eScience and data, the value-added elements made by libraries and librarians will be what contributes to the success of these partnerships. As Dr. Liz Lyon stated at the ARL/CNI Forum numerous times, librarians have two choices: "Transition or Transform."[32] The same applies to the data side of the house.

Serving the needs of the scientific community is never easy but libraries if they want to "play" have to bring new skills and services to the data side of the house.

- Systematically manage and make accessible information from heterogeneous sources – Librarians already know and are engaged in:
  - **Metadata, discovery mechanisms, portals, VRE, "Scientific World"**
  - **Publication and Citation**
  - **Selection and use of tools and resources**
  - **Digitization of legacy content**
  - **Access management, copyright, IPR, Licenses**
  - **Curation and preservation advice**
  - Providing specialist assistance to end-users
  - **Expertise in user services and training**
  - Exploiting strengths in design and implementing innovative and useful e-Science information infrastructure.
- But they must reduce the risk of "re-inventing the wheel" and instead
- Inform data management plans
- Data documentation best practices
- Deliver new integrated support services
- Migrate from library establishment to informatics practice

#45 They must "reinvent the Library" and what this constitutes for the collections side is that the traditional science resources of journals, conference papers & proceedings, government information, books, grey literature and other forms will be heavily influenced by primary data collected in many ways that need to be retained as part of the scholarly enterprise. The emphasis will be more on collecting the front-end of the research than only the final output of the conclusions and the package.

It is going to be an interesting and challenging ride. But we can't be anything but humbled when we think of the breakthroughs alone in 2008 of the advancements in all sciences, but particularly in computational science, computational thinking, TeraGrid computing, and Cloud Computing.[33] 2009 is the 200th birthday of Darwin, the Year of Science, the 125th anniversary of the IEEE and as we engage in those celebrations and milestones, we have only our future to be even more excited about.

As Vannevar Bush wrote in 1945, "Progress…depends upon a flow of new knowledge. New products, new industries and more jobs require continuous additions to knowledge of the laws of nature and the application of that knowledge to practical purposes." To promote eScience and science librarianship, within the realm of grey literature at a time of institutional and scientific competitiveness and economic uncertainty is no small challenge, but an urgent call to action in order to achieve success with managing data.

#46– probably should have inserted another wordle to show how fast things change!

Other Resources

Association of Research Libraries, "To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering," Association of Research Libraries, 2006. http://www.arl.org/pp/access/nsfworkshop.shtml

Association of Research Libraries, ARL Scholarly Communication Steering Committee, *Agenda for Developing E-Science in Research Libraries.* Washington, DC: Association of Research Libraries, 2007.

Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations. Final Repor t from Workshops on Building Effective Virtual Organizations. Washington, DC: National Science Foundation (NSF), May 2008.

Borgman, Christine (2005), "Building a Usable Infrastructure for e-Science: An Information Perspective. Presentation delivered at the e-Science All Hands meeting, Nottingham, UK, September 20, 2005. http://nesc.ac.uk/talks/ahm2005keynote1.ppt

"Collaborative e-Science Libraries," entire issue of *International Journal on Digital Libraries*, 7(1), October 2007.

Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Report by the National Academy of Sciences, National Academy of Engineering and Institute of Medicine of the National Academies. Washington, DC: National Academies Press, 2009.

Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge. A 21st Century Agenda for the National Science Foundation, Report of the NSF Task Force on Cyberlearning. Washington, DC: NSF, June 24, 2008.

Henty, Margaret, "Developing the Capability and Skills to Support eResearch," *Ariadne*, 55, April 2008. Http://www.ariadne.ac.uk/issue55/henty/

Hey, Tony and Trefethen, Anne, "Cyberinfrastructure for e-Science," *Science*, 308, (5723), May 6, 2005.

Hey, Tony and Trefethen, Anne, "The Data Deluge: An e-Science Perspective," in *Grid Computing: Making the Global Infrastructure a Reality*. Chichester, UK: Wiley, 2003:89-824.

Kroski, Ellysa, *Web 2.0 for Librarians and Information Professionals*. New York: Neal-Schuman, 2008.

Lynch, Clifford, "The Shape of the Scientific Article in the Developing Cyberinfrastructure," CT Watch Quarterly, August 2007. http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/

National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," Washington, DC: National Science Foundation, 2005. http://www.nsf.gov/pubs/2005/nsb540/start.jsp

National Science Foundation Office of Cyberinfrastructure – http://www.nsf.gov/oci

Pagano, Pasquale, et al, "Supporting eScience Communities: TheD4Science Perspective on EGI," Presentations made at D4Science Meeting, Paris, France, December 17, 2008. URL>>

Payette, Sandy, ""Fedora Commons Overview and Background," UK Fedora Training, London, UK, January 22-23, 2009. URL: http://74.125.155.132/search?q=cache:G4HoHBF3QpcJ:www.rsp.ac.uk/events/FedoraDay/1-fedora-background.ppt+sandy+payette+AND+%22Fedora+commons+overview+and+background%22&cd=3&hl=en&ct=clnk&gl=us (cited Powerpoint presentation adopted from September 2007 training presentation noted here)

Simpson, Pauline, "Libraries Supporting eScience: Combining Cultures," Presentation delivered at *Digital Libraries a la Carte Workshop*, University of Tilburg, Netherlands, August 27-31, 2007.

Sustaining the Digital Investment: Issues and Challenges of Economically sustainable Digital Preservation. Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Washington, DC: NSF, December, 2008.

Terras, Melissa, "E-Science in the Library and Information Studies Sector: Overview of Seminar Findings for the e-Science Scoping Study." URL>>>>

U.K. Research Council, "E-Science." URL>>>>

---

[1] Julia Gelfand (2009), "The Changing Collections for eScience: What it Means for Libraries," Presentation at the IFLA Section on Acquisitions & Collection Development, Mid-Year Meeting, China Academy of Sciences, Beijing, 26 February 2009.

[2] Malcolm Atkinson, eScience Envoy. See http://www.rcuk.ac.uk/escience/default.htm

[3] JR Balser, A. Barucin, (2008), "Science at the Interstices: An Evolution in the Academy," *Academic Medicine*, 83 99):827-31.

[4] Ibid.

[5] Michael R. Nelson (2009), "The Cloud, The Grid, and the Internet of Things: Potential and Policy," Presentation made at the American Association for the Advancement of Science Meeting, Chicago, IL., February 14,2009.

[6] ICPSR – see http://www.icpsr.umich.edu)

[7] Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, National Academy of Sciences, *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. Washington, DC: National Academy Press, 2009. See http://www.nap.edu/catalog.php?record_id=12615
[8] *Ibid*, 3.
[9] *Ibid*, 3-4.
[10] Alexander M. Bickel, "Passion and patience: Centennial Year Thoughts on the 'Brandeis Way,'" *New Republic*, 12 November 1956, 16.
[11] Dominic J. Farace, "Forward."

[12] See Keith Jeffrey, et al, "CRIS, Grey Literature and the Knowledge Society," 2000. ftp://ftp.cordis.lu/pub/cris2000/docs/jeffery_full and EllyDijk, et al, "Accessing Grey Literature in an Integrated Environment of Scientific Research Information," *Conference Papers of the International Conference on Grey Literature*, 2008, Vol. 9: 15-22.
[13] *Cyberinfrastructure Vision for 21st Century Discovery* (2007). Washington, DC: National Science Foundation, Cyberinfrastructure Council:i.
[14] Ibid: 1-3.
[15] ARL/CNI Forum, "Reinventing Science Librarianship." Arlington, VA: October 17-18, 2008. URL>>>
[16] Chris Anderson (2008), *Wired Magazin*e, June 23, 2008.
[17] Tony Hey, CNI Presentation, Washington, DC, December 2005.
[18] Ibid.
[19] Chris Greer (2008), "E-Science Trends: Transformations & Responses" ARL & CNI Forum on Re-inventing Science Librarianship: Models for the Future, (From here-on cited as ARL/CNI Forum) Arlington, VA: October 26, 2008.
[20] Harold Varmus (2009), The Art and Politics of Science. NY: Norton, 242.
[21] Ibid, 243.
[22] Ibid, 268.
[23] See A World of Science in the Developing World, (2008*). Nature TWAS Supplement*. http://www.nature.com/twas/

[24] HA Piwowar, MJ Becich, H. Bilofsky, RS Crowley (2008), Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers," *PLoS Medicine* 5(9):3183:1-5.
[25] National Center for Research Resources – Strategic Plan 2009-2013. See http://ncrr.nih.gov/strategic_plan/online_version/initiatives.asp

[26] Sayeed Choudhury (2008), "Data Curation Issues and Challenges," ARL/CNI Forum, Arlington, VA: October18, 2008.
[27] Tony Hey (2009), "eScience and Data-Intensive Science," Presentation made at the AAAS Conference, Chicago, IL, February 16, 2009.
[28] Nelson, Slides 8 & 13.
[29] Tony Hey and Jessie Hey (2006), "e-Science and Its Implications for the Library Community," *Library Hi-Tech* 21 (4): 515-528.
[30] Catherine Blake (2008), "Reinventing Science Librarianship: Education for New Roles," Presentation at ARL/CNI Forum, Arlington, VA, February 18, 2008.
[31] See http://lis.uiuc.edu/programs/cpd/DC_Inst/

[32] Liz Lyon, (2008), "Transition or Transform: Repositioning the Library for the Petabyte Era," ARL/CNI Forum, Arlington, VA: October 18, 2008.
[33] *Report of the Panel on Recent Significant Advancements in Computational Science Breakthroughs 2008.* Washington, DC: US Department of Energy, Office of Science; and *TeraGrid 2008 Science Highlights*. Washington, DC: National Science Foundation's Office of Cyberinfrastructure, 2008. Also, see Bruce Alberts, "Editorial," *Science* 322, December 19, 2008. DOI: 10.1126/science.116962.