e-Science, Data Curation and the Role of the Information Specialist:

An Annotated Bibliography

Helen Helm
INFO 522-900: Information Access & Resources
March 16, 2012
Introduction and Scope

The following bibliography brings together peer-reviewed journal articles which discuss the curation of data sets produced by e-science. The articles, published between 2003 and 2012, touch upon the history of scientific research; how it has changed with the advent of supercomputing, the World Wide Web, and social media; and the problem with documenting, archiving and making available this vast collection of research output. Scholars at the front of this challenge describe their current understanding of the problem and the particular contributions librarians can make to data curation projects. The articles report the results of projects undertaken both in the United States and the United Kingdom to create the workflows, technological infrastructure, and the taxonomy useful for archiving and making available the products of e-science. Since the literature on all the aspects of e-science is so vast, this bibliography narrows the topic to issues surrounding data curation and the role libraries and information specialists play in the resolution of the data deluge crisis.

Description

E-Science and the vast amounts of data it produces is changing the way that research is undertaken and reported. In the old days, researchers ran experiments, wrote up the conclusions and results in lab notebooks, and reported their work by publishing articles in peer-reviewed journals. Now, however, “new science,” as it is named by Joint, consists of huge, costly projects generating terabytes of data, all of which cannot be exhausted of its information in a few journal articles (Joint, 2007, p. 452). Many people are looking at this raw data as a goldmine of information some of which is being lost to scholarship because there is not a standard mechanism by which to archive the data for future use. Libraries and librarians are in a position to contribute their skills to solve the problem of saving the valuable e-science materials for future generations.

Summary of Findings

e-Science is a new phenomenon made possible by the advances in computer technology which have occurred in the past ten years or so. With the development of the World Wide Web, distributed computing, social media, and Web 2.0, e-science has come into its own. But, what is
e-science? Several definitions appear in the literature. According to Hey and Trefethen (2003), “e-science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it” (p. 1809). Chen (2007) in her article qualified this definition adding that this infrastructure enables scientists accomplish more in less time (p. 313). e-Science has the ability to bring many researchers and resources together to pursue research goals which individual institutions or researchers could never do on their own. Voss (2009) described this same phenomenon as e-research (p. 175). A famous example of this type of e-science project is the Large Hadron Collider (LHC) experiments being done at the CERN (Conseil Européen pour la Recherche Nucléaire) in Geneva, Switzerland.

They great ability of e-science to produce more information contributing to the collective knowledge of the human race is certainly an advancement for scholarship. All this data, however, brings its own troubles to the scholarly world that were never there before. To collect, catalog, and make available the scholarship of former years, libraries and librarians worked with the end product of research – the published journal article. Now, however, with the advent of e-science, researchers are finding out that more than the published journal article should be saved.

The vast amount of data generated by many sophisticated instruments becomes valuable in its own right. Choudhury (2008) put it well when he wrote “scientists wish to produce new forms of publications that comprise both articles and data, both of which can be traced back to source data in distributed repositories or content stores” (p. 215). Joint (2007) found that large-scale experiments could not be repeated if the data was needed again, “so the data itself [was] the most important outcome of the experiment” (p. 452). In the same vein, Heidhorn (2008) observed that reusing the data by other research projects saved money by saving other researchers from having to produce the same data all over again (p. 290). Funding agencies are also beginning to realize that the data generated by the projects they fund are truly the most valuable part of the project. In 2012, as reported by Heidorn (2011), the “National Science Foundation began requiring data management plans with most grant proposals” (p. 665).

Putting this data in a form that can be archived and made searchable for reuse, however, proves to be a difficult undertaking. First, the amount of data itself can be overwhelming. Wong (2009), Heidorn (2011), Hey & Trefethen (2003), and Joint (2007), all reported on the amazing amount of data that comes out of e-science research projects. Wong (2009) listed genomics,
climate research, high energy physics and astronomy” as the areas which produce the most data (p. 125).

Second, the type of data has changed. The data revealing new information about nature cannot be observed like an ant or a planet. Joint (2007) put it well in this description of a presentation he saw given by a researcher from CERN:

Rather than being ‘determinant entities’ that can be stuffed and mounted in a museum, fundamental particles are ‘information events’ – slide 6 of Hoffman’s CERN presentation … is his nearest attempt to photographing a fundamental particle. What we see is a grid with numbers on it, some of which are circled. There is an arrow pointing to the circle saying ‘Interesting physics!’ What is inside the circle is a data pattern, not a ‘thing’ as such – the particle that generated the data pattern died more or less as soon as it was born. In fundamental physics, when you pull your nets in, the creature of the depths has disappeared! (p. 453)

Third, the multiple formats in which the data is stored, not only between different projects, but even in the same project makes building a repository which can accommodate them technically difficult. Zimmerman in her study pointed out that when the data sets are small and very diverse, it is difficult to describe them adequately so that others can retrieve them from the repository (2007, p. 7). Lage, Losoff, & Maness (2011) described researchers who did have plans for managing their primary data but did not know what to do with other data sets which they had collected, of minor interest to their particular project, yet still valuable. Either the orphan data was of a different type from the primary data, or the format was a different type which did not fall into the maintenance plans devised (p. 922). Purdue is studying this situation with a project named “e-Scholar.” To solve the format problem, e-Scholar links three databases together, one for archives, one for documents, and one for the raw datasets instead of trying to put the different formats into the same database (Witt, 2008, p. 196).

A fourth problem encountered by data curation projects is the question of intellectual property rights. Who owns the data produced by a multi-institution research
project? Who will control access to the data? As Carlson (2012) put it “intellectual property rights and protections is another subject that presents a set of potentially thorny issues that may affect a researcher’s willingness or ability to share their data set” (p. 18). Zimmerman (2007) also found unwillingness for scientists to spend the time or resources preserving their data because they did not see any reward for sharing the data: all the reward was in publishing journal articles, not the data sets (p. 13).

The fifth concern proves to be the most pressing, however: the lack of standardized metadata for describing the data sets making them accessible and, therefore, more valuable. Garritano & Carlson (2009) described the situation best: “the metadata used to organize the data will … play a key role in how that data will be discovered and repurposed, especially by those unfamiliar with the initial project” (p. 50). Hey (2003) realized that each research community, in order to design search engines that work, needed “to come together to define generally accepted metadata standards for their community data grids” (p. 1820). Wong (2009) also lamented a metadata standard writing that “while citation standards for traditional research publications are well-established, a robust data citation standard is not yet in place” (p. 130).

With these concerns in mind, many researchers are looking to the library science community as necessary collaborators in finding viable solutions to these problems. Garritano (2009) reported the value of having librarians collaborate on a project in order to win funding from the NSF (p. 4). Quite a few researchers welcomed the abilities of research librarians to work across disciplines and form liaisons with different groups necessary for the success of their collaborative e-science projects (Carlson, 2012, p. 10), (Kesselman, 2009, pp. 392-393), and (Carlson, 2009, p. 4). Carlson (2012) also observed that reference librarians could “assume the role of a trusted data consultant” by helping researchers form liaisons with those who understand the technology needed for curating their research and helping them to understand the issues involved (p. 12). Hey (2006) pointed out the value of the librarians’ ability to analyze ideas from a broad perspective and their ability to create good citations for data sets (p. 525). He then made the observation that “the e-science revolution will put libraries and repositories center stage in the development of the next generation research infrastructure” (p. 526).
Acquiring the skills necessary for undertaking this data curation can appear quite daunting to the information professional. As Heidorn (2011) aptly reported it, "data require different structural metadata, schemas, and vocabularies. Librarians who have adapted their skills are difficult to find" (p. 670). So what should an information scientist who is interested in working on an e-science project do to gain these skills? United States Universities in the United States who are venturing out into the teaching of e-science information professionals include the University of Illinois, the University of North Carolina, and the University of Arizona (Heidorn, 2011, p. 294). The Purdue Libraries have produced the Data Curation Profile Toolkit (DCP Toolkit) to help train librarians wishing to learn how to conduct a data interview (Carlson, 2012, p. 8). Other resources information specialists can consult are the UK Digital Curation Centre (DCC) (Heidorn, 2011, p. 665), the National Science Foundation’s Sustainable Digital Data Preservation and Access Network Partners (Heidorn, 2008, p. 294), and Purdue University’s Distributed Data Curation Center (D2C2) (Witt, 2008, p. 197).

What will happen to research and the benefits it provides to society if information specialists do not rise to the occasion? Osswald (2007) stated it well when he said others will provide the liaison services, apply metadata to the data sets and archive the research; however, it will be done inadequately. He also warned that “library and information services (LIS) institutions might lose their connection to scientists” (p. 517). He also doubted that other professionals will appear to fill the gap (p. 522). Heidorn (2011) warned that some other institution will curate the data sets while publishers will curate the articles associated with the data. In this case, the data will be separated from the articles written to interpret the data, a less than optimum solution (p. 670). Joint (2007) put it best when he said “for the information society to work, information must be held in common. Librarians facilitate this by designing the information structures that underpin the process of knowledge exchange, and they can do this as much for the new science as they have always done for the old" (p. 454).

Bibliography
Entry 1:


Abstract:

Purpose – As libraries become more involved in curating research data, reference librarians will need to be trained in conducting data interviews with researchers to better understand their data and associated needs. This article seeks to identify and provide definitions for the basic terms and concepts of data curation for librarians to properly frame and carry out a data interview using the Data Curation Profiles (DCP) Toolkit.

Design/methodology/approach – The DCP Toolkit is a semi-structured interview designed to assist librarians in identifying the data curation needs of researchers. The components of the DCP Toolkit were analyzed to determine the base level of knowledge needed for librarians to conduct effective data interviews. Specific concepts, definitions, and examples were sought through a review of articles, case studies, practitioner resources and from the experiences of the Purdue University Libraries.

Findings – Data curation concepts and terminology are not yet well-defined and often vary across, or even within fields of study. This research informed the development of a workshop to train librarians in using the DCP Toolkit. The definitions and concepts addressed in the workshop include: data, data set, data lifecycle, data curation, data sharing, and roles for reference librarians.

Practical implications – Conducting a data interview can be a daunting task given the complexity of data curation and the lack of shared definitions. Practical tools and training are needed to help librarians develop capacity in data curation.

Originality/value – This article provides practical information for public service librarians to help them conceptualize and conduct a data interview with researchers.

KeywordsTerminology, Interviews, Librarianship, Training, Roles
**Paper type** Case study

**Annotation:** Working in the Purdue University Libraries/Distributed Data Curation Center, one of the major centers for research in data curation, the author plays an important role in the ongoing discussion on data curation. He discusses the necessity for developing a toolkit to help reference librarians conduct a data interview with researchers. Reading this article will give librarians a tool by which to learn the research methods and sharing patterns used by researchers.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science and science fields. I used a keyword approach to finding this article. After locating the article in Scopus, I did a title search in Summon on the Hagerty Library website. I accessed the article from the Emerald website.

**Database:** SciVerse Scopus

**Method of Searching:** Keyword [executed 3/2/2012]

**Search String:** ((escience OR e-science OR "data curation") AND librarian)

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with a content type of academic/scholarly. The journal website describes this journal as refereed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 2:**


**Abstract:**

At Johns Hopkins University, the institutional repository (IR) is being developed as a component of an overall digital library architecture that will emphasize long-term preservation. The IR represents a set of services that will be developed to support the identified needs or requirements of faculty and students. Given the research-intensive environment at Johns Hopkins, one particular area of interest relates to data sets from a diversity of disciplines ranging from the humanities to
the sciences. Essentially, the IR is being developed as a “gateway” to the underlying digital archive that will support data curation as part of an evolving cyberinfrastructure featuring open, modular components. In addition to this technological framework, Johns Hopkins is developing new roles and relationships between the library and the academic community, most notably through the development of “data scientists” or “data humanists.” These developments reflect the realization that the IR is the first step in a longer journey and that for institutional efforts to be successful, they must be integrated into a larger landscape of repositories that serve a distributed and diverse academic community.

**Annotation:** Choudhury makes the point that some institutions viewed institutional repositories more as a way to make themselves famous rather than seeing what it should be - an aid to their scholars. As a writer of numerous articles and principle investigator on high-profile projects, the author is well-versed in the topic of data curation. He adds to the literature by reporting on the project for developing the Institutional Repository at Johns Hopkins University.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating the article, I did a title search in Summon on the Hagerty Library website then accessed the article through ProQuest.

**Database:** Library Literature and Information Science [Dialog]

**Method of Searching:** Keyword search [executed 2/24/2012]

**Search String:**

```
S3 DATA()CURATI?
S4 S3 AND LIBRAR?
```

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with an academic/scholarly content. The journal website reports that each article is peer-reviewed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.
Entry 3:


Abstract:

The challenge of digital preservation of scientific data lies in the need to preserve not only the dataset itself but also the ability it has to deliver knowledge to a future user community. A true scientific research asset allows future users to reanalyze the data within new contexts. Thus, in order to carry out meaningful preservation we need to ensure that future users are equipped with the necessary information to re-use the data. This paper presents an overview of a preservation analysis methodology which was developed in response to that need on the CASPAR and Digital Curation Centre SCARP projects. We intend to place it in relation to other digital preservation practices, discussing how they can interact to provide archives caring for scientific data sets with the full arsenal of tools and techniques necessary to rise to this challenge.

Annotation: This article adds significantly to the literature on designing a model for data preservation. Similar to the article by Carlson, (2012), it describes a method by which to plan digital preservation projects. Its depth, however, is much greater than the latter since it proposes a model for preserving not only the data but also the relevant information about the data as well as the software needed to utilize it. The authors, as members of the Science and Technology Facilities Council, UK, are highly cited in their field.

Search Strategy: I chose this web resource because it indexes journal articles and other resources authored by university faculty and researchers. I used a keyword search with the terms data curation and found *The International Journal of Digital Curation*. I followed the link to the current issue. There, I browsed through the titles, found the article in the peer-reviewed section of the journal and downloaded it from the journal website.

Database: INFOMINE

Method of Searching: Browsing [executed 3/14/2012]
Search String: N/A

Scholarly/Refereed Status: Ulrich’s lists this journal as academic/scholarly and indexed as reviewed. The journal website describes this journal as scholarly and peer-reviewed. All of the articles in the journal are not peer-reviewed, but this particular article is marked as being peer-reviewed. LISA does not index this journal. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

Entry 4:


Abstract:
For liaison or subject librarians, entering into the emerging area of providing researchers with data services or partnering with them on cyberinfrastructure projects can be a daunting task. This article will provide some advice as to what to expect and how providing data services can be folded into other liaison duties. New skills for librarians and traditional skills that can be adapted to data curation work will also be discussed. A case study on the authors' experiences collaborating with two chemistry faculty on an e-science project serves as the framework for the majority of this article.

Annotation: Unlike the Heidorn (2011) article, this article does not discuss so much the technical skills for data curation. Instead, Garritano, of Purdue University, discusses how subject librarians can contribute by acting as liaisons between different parties in an inter-institutional research project. He adds to the literature by describing the CASPiE project at Purdue University. Of particular interest is his discussion of the librarian’s role in the grant proposal process and the ways librarians can add value to the proposed project.

Search Strategy: I chose this database because it indexes important journal articles in the information science and computer science fields. I used a keyword approach to finding this article. After locating the citation, I did a title search in Summon on the Hagerty Library
website. I followed the link given in the library record to the journal website. I browsed through the issues to get to this article.

**Database:** INSPEC [Dialog]

**Method of Searching:** Keyword search [executed 2/24/2012]

**Search String:**

\[
S1 \text{ ESCIENCE? OR E()SCIENCE?}
\]

\[
S2 \text{ S1 AND LIBRAR?}
\]

**Scholarly/Refereed Status:** Ulrich’s lists this journal as a peer-reviewed, scholarly/academic journal. All of the articles in this journal are not peer-reviewed; however, the journal website attached a peer-reviewed icon next to this article indicating that this particular article is peer-reviewed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 5:**


**Abstract:**

One of the primary outputs of the scientific enterprise is data, but many institutions such as libraries that are charged with preserving and disseminating scholarly output have largely ignored this form of documentation of scholarly activity. This paper focuses on a particularly troublesome class of data, termed dark data. “Dark data” is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost. The article discusses how the concepts from long-tail economics can be used to understand potential solutions for better curation of this data. The paper describes why this data is critical to scientific progress, some of the properties of this data, as well as some social and technical barriers to proper management of this class of data. Many potentially useful institutional, social, and technical solutions are under development and are
introduced in the last sections of the paper, but these solutions are largely unproven and require additional research and development.

**Annotation:** Written by a highly-cited author, this article is valuable in that it discusses barriers to data curation. He describes types of data not easily found and points out the importance of saving them. This article is also valuable for its discussion of the possible solutions to the barriers to data curation.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating the article in Web of Science, I followed the “Get It” button under the record. It took me to the Hagerty Library record where I accessed the article from Project MUSE.

**Database:** Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index [Web of Science]

**Method of Searching:** Keyword [executed 2/27/2012]

**Search String:** Topic=((escience or e-science or data curation) and librar*) Refined by: Web of Science Categories=( INFORMATION SCIENCE LIBRARY SCIENCE ) Timespan=All Years. Lemmatization=On

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with an academic/scholarly content. The journal website reports that each article is peer-reviewed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 6:**


**Abstract:**

The role of libraries is to collect, preserve, and disseminate the intellectual output of the society. This output includes books and serials as well as the digital
versions of the same. Scientists, other scholars, and all of society are now producing, storing, and disseminating digital data that underpin the aforementioned documents in much larger volumes than the text. The survival of this data is in question since the data are not housed in long-lived institutions such as libraries. This situation threatens the underlying principles of scientific replicability since in many cases data cannot readily be collected again. Libraries are the institutions that could best manage this intellectual output.

Annotation: This article is important for giving an overview and explaining the role of libraries in the curation of research data. The author, P.B. Heidorn, writes from the University of Arizona, one of the first universities to start a digital curation program. His contribution is a thorough description of the issues surrounding data curation and its importance to the advancement of scholarly endeavors.

Search Strategy: I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After finding the citation, I did a title search in Summon on the Hagerty Library website and followed the link given by the library. I accessed the article from InformaWorld.

Database: Library Literature and Information Science [Dialog]

Method of Searching: Keyword search [executed 2/24/2012]

Search String: S1 ESCIENCE? OR E()SCIENCE?
S2 S1 AND LIBRAR?

Scholarly/Refereed Status: Ulrich’s lists this journal as refereed and described as having academic/scholarly content. The journal website reports that all the articles in this journal are screened by the editor and peer-reviewed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Abstract:**

**Purpose** – The purpose of this article is to explain the nature of the “e-Science” revolution in twenty-first century scientific research and its consequences for the library community.

**Design/methodology/approach** – The concepts of e-Science are illustrated by a discussion of the CombeChem, eBank and SmartTea projects. The issue of open access is then discussed with reference to arXiv, PubMed Central and EPrints. The challenges these trends present to the library community are discussed in the context of the TARDis project and the University of Southampton Research Repository.

**Findings** – Increasingly academics will need to collaborate in multidisciplinary teams distributed across several sites in order to address the next generation of scientific problems. In addition, new high-throughput devices, high-resolution surveys and sensor networks will result in an increase in scientific data collected by several orders of magnitude. To analyze, federate and mine this data will require collaboration between scientists and computer scientists; to organize, curate and preserve this data will require collaboration between scientists and librarians. A vital part of the developing research infrastructure will be digital repositories containing both publications and data.

**Originality/value** – The paper provides a synthesis of e-Science concepts, the question of open access to the results of scientific research, and a changing attitude towards academic publishing and communication. The paper offers a new perspective on coming demands on the library and is of special interest to librarians with strategic tasks.

**Annotation:** This article is interesting in that it gives a review of the history of the partnership between distributed computing and collaborative science. The authors, highly involved in the computer science field, discuss the importance of good metadata and look to the library community for the skills necessary for applying metadata to the data sets.
Search Strategy: I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating this citation, I did a title search in Summon on the Hagerty Library website. I accessed the article from ProQuest.

Database: Library Literature & Information Science [Dialog]

Method of Searching: Keyword search [executed 2/24/2012]

Search String: S1 ESCIENCE? OR E()SCIENCE?
S2 S1 AND LIBRAR?

Scholarly/Refereed Status: Ulrich’s indexes this journal as refereed with a content described as academic/scholarly. The journal website explains that an article is reviewed by the editor and, if approved, sent to two referees for review using a double blind review method. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

Entry 8:


Abstract:

After a definition of e-science and the Grid, the paper begins with an overview of the technological context of Grid developments. NASA's Information Power Grid is described as an early example of a 'prototype production Grid'. The discussion of e-science and the Grid is then set in the context of the UK e-Science Programme and is illustrated with reference to some UK e-science projects in science, engineering and medicine. The Open Standards approach to Grid middleware adopted by the community in the Global Grid Forum is described and compared with community based standardization processes used for the Internet, MPI, Linux and the Web. Some implications of the imminent data deluge that will arise from the new generation of e-science experiments in terms of archiving and
curation are then considered. The paper concludes with remarks about social and technological issues posed by Grid-enabled 'collaboratories' in both scientific and commercial contexts.

**Keywords**: e-Science; Grid; data curation; data preservation; open standards

**Annotation**: After reading this article, one will gain a good overview of the e-science phenomenon. The author, having written many articles on the subject of information, knowledge and technology, again contributes to the literature by discussing projects current in 2003. Especially interesting is his explanation of how persons from the library science field who are expert in metadata assignment will be of utmost importance to the success of the e-science endeavor.

**Search Strategy**: I chose INSPEC because it indexes important journal articles in the library and information science and computer science fields. I used a keyword approach to finding this article. Since this database covers the electronic engineering and computer engineering fields, I found it not useful to use the keyword “library” as it retrieves many articles not associated with the library as a repository for books. After locating the article in INSPEC, I clicked the “full text” button then accessed the article from the journal website.

**Database**: INSPEC [Engineering Village]

**Method of Searching**: Keyword [executed 2/11/2012]

**Search String**: e-science AND data curation in all fields

**Scholarly/Refereed Status**: Ulrich’s indexes this journal as refereed with a content type of academic/scholarly. The journal website describes this journal as peer-reviewed. LISA does not index this journal. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 9**:


**Abstract**:
**Purpose** – This paper aims to outline the information management principles of the so-called “new science”, and to attempt to put these in the context of traditional library and information science principles.

**Design/methodology/approach** – A brief review of some work in the area, in particular focusing on the work showcased by the annual digital preservation conference series hosted by the Digital Curation Centre in Scotland (www.dcc.ac.uk/).

**Findings** – There is a danger that scientists (as opposed to LIS professionals) will apply the information management techniques of the new science to their own activities inappropriately, especially to research that is best curated as “old” not new science. This is something on which information professionals are well placed to give advice and make judgements.

**Research limitations/implications** – More practice-oriented research is needed to enhance understanding of how traditional librarianship practices can be applied to the data-intensive scientific research carried out by so-called “virtual organisations”.

**Practical implications** – This paper makes some initial suggestions about the how the tools of library and information practice can be related to the “new science”. In particular, it highlights their relevance to distinguishing between the information management needs of the “old” and the “new” sciences: these needs are quite distinct, though easily confused.

**Originality/value** – This paper relates terms from pure science such as the virtual organisation, cyberinfrastructure, and e-Science to traditional LIS concepts, and tries to create an understanding of the relationship between the two disciplines for the library practitioner.

**Keywords** Libraries, Digital libraries, Data handling, Archives management

**Paper type** General review

**Annotation**: The issues discussed in this article are far-reaching and highly important to the advancement of learning and research. Adding to the body of literature on data curation, the author discusses the necessity of making the results of the “new science,” as he calls it, available for dissemination to the whole of society. The practitioner librarian must have a say in the
curation of the data in order to make this type of knowledge open to other disciplines. Since the author works at the Centre for Digital Library Research, Glasgow, UK, he is quite competent to speak on this subject.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating this citation, I did a title search in Summon on the Hagerty Library website. I then accessed the article through ProQuest.

**Database:** SciVerse Scopus

**Method of Searching:** Keyword [executed 3/2/2012]

**Search String:** ((escience OR e-science OR "data curation") AND librarian)

**Scholarly/Refereed Status:** Ulrich’s lists this journal as refereed with content type of academic/scholarly. The journal website describes this journal as reviewed by the editor and one or more reviewers, double-blind. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 10:**


**Abstract:**

Embedded librarianship focuses on the user and brings the library and the librarian to the user, wherever they are – office, laboratory, home, or even on their mobile device. This article provides an overview of the various ways libraries and librarians are embedding themselves into research and learning environs. Several roles are highlighted, including course-integrated instruction librarians as members of research teams, librarians collaborating with faculty in scholarly communication activities and librarians as partners in multidisciplinary, global, and virtual collaborations. Definitions of key terms precede the overview and
provide context; consideration of the human resources side of the equation follows. Reflections on organizational structure conclude the article.

**Annotation:** The authors, both working in well-known academic libraries, discuss the different scenarios in which librarians can “embed” themselves into research teams and academic learning. Although not strong in technical information, this article does describe the many situations where an embedded librarian can contribute to individual research teams, including research teams engaged in e-science. Especially interesting is their discussion on the different skill sets that librarians bring to the e-science endeavor.

**Search Strategy:** I chose this database I wanted to see what Google Scholar has on the subject. Google Scholar indexes scholarly web content. I used a keyword approach and found the first article using these search terms: librar* escience OR “e science”. The article itself is not peer-reviewed; however, I found this citation in the list of references. After locating the citation, I did a title search in Summon on the Hagerty Library website. I accessed the article from Taylor & Francis Online.

**Database:** Google Scholar

**Method of Searching:** Footnote chasing [executed 2/5/12]

**Search String:** Referenced in:


**Scholarly/Refereed Status:** Ulrich’s lists this journal as refereed with a content type as academic/scholarly. The journal website states that all papers published undergo screening and peer review. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 11:**

**Abstract:**
Increasingly libraries are expected to play a role in scientific data curation initiatives, i.e., “the management and preservation of digital data over the long-term.” This case study offers a novel approach for identifying researchers who are receptive toward library involvement in data curation. The authors interviewed researchers at the University of Colorado Boulder and, after analysis, created eight design “ personas.” Each persona represents an aggregation of researcher attributes and can be used to target strategic relationships for nascent or emerging data management initiatives. These personas are applicable to any academic library seeking to provide data curation support.

**Annotation:** Creating personas for the purpose of studying scientific data curation is a new use of this model for defining user needs. The authors, who participated in the study done by the University of Colorado Boulder, report using a persona model in the unique setting of scientific data curation. This paper contributes to the literature with its novel approach to the data interview.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating the citation, I searched for the publication title in the e-journals on the Hagerty Library website and navigated to Project Muse. I located the article by doing a title search in Project MUSE.

**Database:** Library Literature and Information Science [Dialog]

**Method of Searching:** Keyword search [executed 2/24/2012]

**Search String:** S3 DATA CURATI?
S4 S3 AND LIBRAR?
Scholarly/Refereed Status: Ulrich’s lists this journal as refereed with a content type of academic/scholarly. The journal website describes this journal includes peer-reviewed articles. This article is one of the peer-reviewed articles. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

Entry 12:


Abstract:

**Purpose** – This paper aims to analyse whether, and how far, library and information services (LIS) are involved in e-Science and grid computing projects funded by authorities in the European Union and Germany. It explains and emphasises the relevance of LIS-based information services and expertise in e-Science activities.

**Design/methodology/approach** – Projects were analysed to determine the extent to which institutions or persons with LIS expertise were involved and information services provided.

**Findings** – Very few of the e-Science projects made use of LIS-based services. This omission could reduce the quality of e-Science-related research in the long term and LIS experts could lose their role and expertise in serving scientists. Further research is needed to verify this. Additionally research is needed to determine whether e-Science projects in other countries are also lacking LIS services.

**Research limitations/implications** – The research was undertaken during the preliminary phase of the projects. Practical implications – Funding should emphasise the inclusion and provision of LIS services.

**Originality/value** – The findings show a lack of LIS services in e-Science projects, which could otherwise enhance the projects’ development and the distribution of their results.
Annotation: Writing in 2007, Osswald, a member of the Institute of Information Science in Cologne, Germany, finds the relationship between e-science and library services as woefully lacking. He reports on the numerous e-science projects, especially in Germany, which do not include an LIS-based service. This article is interesting in that it gives an idea of the situation in 2007 of the lack of inclusion of library services in the conduct of e-science and the losses to research because of this lack.

Search Strategy: I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After finding the citation, I did a title search in Summon on the Hagerty Library website and accessed the article from Emerald.

Database: Library Literature and Information Science [Dialog]

Method of Searching: Keyword search [executed 2/24/2012]

Search String: $S1$ SCIENCE? OR $E()SCIENCE$?  
$S2$ $S1$ AND LIBRAR?

Scholarly/Refereed Status: Ulrich’s describes this journal as having academic/scholarly content and as being refereed. The journal website also describes this journal as refereed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.
Entry 13:


Abstract:

Large, collaboratively managed datasets have become essential to many scientific and engineering endeavors, and their management has increased the need for "eScience Professionals" who extend librarianship into solving large scale information management problems for researchers and engineers. This article focuses on understanding the dimensions of work, worker, and workplace, including the knowledge, skills, and abilities needed for eScience professionals. We used focus groups and interviews to explore the needs of scientific researchers and how these needs may translate into curricular and program development choices. A cohort of five master's students also worked in targeted internship settings and completed internship logs. Results showed that students worked in three major areas: data management, communications between technical and non-technical staff, and science-related functions. We organized this evidence into a job analysis that can be used for curriculum and program development at schools of information and library science. We conclude with suggestions that the emerging eScience profession comprises a promising educational and research focus for information and library science in the coming decade and that science and R&D labs are an underappreciated setting for productive librarianship.

Annotation: The authors, all members of the School of Information at Syracuse University, propose a mixture of skills coming from both the information science field and the regular science field as a training approach for e-science professionals. To research their ideas they conducted a study of the knowledge, skills, and abilities needed on the job for the curation of e-science data. This article contributes to the literature by reporting on the study, its methods, and the results of their findings, including suggestions of courses for training e-science professionals.
Search Strategy: I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating this citation, I did a title search in Summon on the Hagerty Library website. I accessed the article through ProQuest.

Database: Library Literature & Information Science [Dialog]

Method of Searching: Keyword search [executed 2/24/2012]

Search String: S1 ESCIENCE? OR E()SCIENCE?
                 S2 S1 AND LIBRAR?

Scholarly/Refereed Status: Ulrich’s lists this journal as refereed with a content described as academic/scholarly. The journal website describes this journal as a scholarly periodical in which all the articles are refereed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

Entry 14:


Abstract:

Purpose - The purpose of this paper is to investigate the implications of the emergence of virtual research environments (VREs) and related e-research tools for scholarly work and communications processes.

Design/methodology/approach - The concepts of VREs and of e-research more generally are introduced and relevant literature is reviewed. On this basis, the authors discuss the developing role they play in research practices across a number of disciplines and how scholarly communication is beginning to evolve in response to the opportunities these new tools open up and the challenges they raise.
**Findings** - Virtual research environments are beginning to change the ways in which researchers go about their work and how they communicate with each other and with other stakeholders such as publishers and service providers. The changes are driven by the changing landscape of data production, curation and (re-)use, by new scientific methods, by changes in technology supply and the increasingly interdisciplinary nature of research in many domains.

**Research limitations/implications** - The paper is based on observations drawn from a number of projects in which the authors are investigating the uptake of advanced ICT in research. The paper describes the role of VREs as enablers of changing research practices and the ways in which they engender changes in scholarly work and communications.

**Practical implications** - Librarians and other information professionals need to be aware of how advanced ICTs are being used by researchers to change the ways they work and communicate. Through their experiences with the integration of virtual learning environments within library information services, they are well placed to inform developments that may well change scholarly communications fundamentally.

**Originality/value** - The paper contributes to emerging discussions about the likely trajectory and impact of advanced ICTs on research and their implications for those, such as librarians and other information professionals, who occupy important support roles.

**Annotation:** The strength of this article is in its description of the various Virtual Research Environment (VRE) software applications that are available. The author describes the different VRE’s used by the different disciplines from bioinformatics to social interaction modeling. Both authors work professionally in the National Centre for E-Social Science, making them qualified to speak on this topic. The article considers the role of information professionals although not in much depth and mostly in passing.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After
locating the citation, I clicked the “Get It” button on Web of Science. I accessed the article on the Emerald website.

**Database:** Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index [Web of Science]

**Method of Searching:** Keyword [executed 2/27/2012]

**Search String:** Topic=((escience or e-science or data curation) and librar*) Refined by: Web of Science Categories=( INFORMATION SCIENCE LIBRARY SCIENCE ) Timespan=All Years. Lemmatization=On

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with a content described as academic/scholarly. The journal website explains that an article is reviewed by the editor and, if approved, sent to two referees for review using a double blind review method. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 15:**


**Abstract:**

Broadly speaking, the lack of a framework for organizing, preserving, and making research data available for the long term has resulted in valuable datasets becoming lost or discarded. The approach of the Distributed Data Curation Center of the Purdue University Libraries has been to integrate librarians and the principles of library and archival sciences with domain sciences, computer and information sciences, and information technology to address the challenges of managing collections of research data and to learn how to better support interdisciplinary research through data curation. One piece of infrastructure that supports these activities is a "distributed institutional repository" that includes electronic documents, digitized archival collections, and research datasets housed
in multiple systems that are connected together using Web Services and other middle-ware. Concurrently, roles for librarians and institutional repositories in data curation are being explored.

**Annotation:** This article is useful for its discussion of the Purdue University data repository and the features of its design. Its contribution to the literature includes a discussion of the techniques used by Purdue for handling the problem of curating different file formats.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science field. I used a keyword approach to finding this article. After locating the citation in Web of Science, I clicked the “Get It” button to get the list of providers. I accessed the article from Project MUSE.

**Database:** Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index [Web of Science]

**Method of Searching:** Keyword [executed 2/27/2012]

**Search String:** Topic=((escience or e-science or data curation) and librar*) Refined by: Web of Science Categories=( INFORMATION SCIENCE LIBRARY SCIENCE ) Timespan=All Years. Lemmatization=On

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with an academic/scholarly content. The journal website reports that each article is peer-reviewed. Searching this journal in LISA, I found it listed in the peer-reviewed journals. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 16:**


**Abstract:**

The profusion of data created by modern research has brought about increasing discussion on the practice of data management and the development of such
infrastructure. Many universities and libraries are exploring how to address this issue using institutional repositories, the Hong Kong University of Science and Technology (HKUST) Library being one of them. As part of the exploration, the author conducted a case study of datasets at major institutional repositories. This paper visits the issues relating to data archiving, reports common practices of data treatment at institutional repositories, and presents a list of recommendations for experimenting data archiving at the HKUST Institutional Repository.

**Annotation:** The strong points evident in this article is the detail with which the author describes the different types of data repositories, the formats encountered, and the types of data files encountered in the repositories she investigated in this study. Especially interesting is her discussion of the software necessary for reading the files and the issue of a citation standard for datasets comparable to that for article publications. Much of the discussion is specific to DSpace, a repository software application, which is the platform used by HKUST. The other articles reviewed do not go into this amount of detail concerning these practical matters: therefore, this is a useful article to read.

**Search Strategy:** I chose this database because it indexes important journal articles in the library and information science and science fields. I used a keyword approach to finding this article. After locating the citation, I clicked the “Full Text” button on the Web of Science record. I accessed the article from ScienceDirect.

**Database:** Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index [Web of Science].

**Method of Searching:** Keyword [executed 2/27/2012]

**Search String:** Topic=((escience or e-science or data curation) and librar*) Refined by: Web of Science Categories=( INFORMATION SCIENCE LIBRARY SCIENCE ) Timespan=All Years. Lemmatization=On

**Scholarly/Refereed Status:** Ulrich’s indexes this journal as refereed with an academic/scholarly content. The journal website describes this journal as a scholarly peer-reviewed journal. Searching this journal in LISA, I found it listed in the peer-reviewed journals.
Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

**Entry 17:**


**Abstract:**

An important set of challenges for eScience initiatives and digital libraries concern the need to provide scientists with the ability to access data from multiple sources. This paper argues that an analysis of scientists’ reuse of data prior to the advent of eScience can illuminate the requirements and design of digital libraries and cyberinfrastructure. As part of a larger study on data sharing and reuse, I investigated the processes by which ecologists locate data that were initially collected by others. Ecological data are unusually complex and present daunting problems of interpretation and analysis that must be considered in the design of cyberinfrastructure. The ecologists that I interviewed found ways to overcome many of these difficulties. One part of my results shows that ecologists use formal and informal knowledge that they have gained through disciplinary training and through their own data-gathering experiences to help them overcome hurdles related to finding, acquiring, and validating data collected by others. A second part of my findings reveals that ecologists rely on formal notions of scientific practice that emphasize objectivity to justify the methods they use to collect data for reuse. I discuss the implications of these findings for digital libraries and eScience initiatives.

**Annotation:** The author discusses the problem of curating many small pieces of data in diverse formats in the same dataset. Because the data is so diverse, they are difficult to describe using only metadata. This article contributes to the literature by examining the social aspect of collecting and sharing data. The author proposes that adding social reaction between the researcher who wants to collect data for reuse and the researcher who collected the data initially will make the data more valuable because it reveals the circumstances surrounding their creation.
Search Strategy: I chose this strategy because, after searching in ArticleFirst with the keyword escience, I found an article by Baru in the *International Journal on Digital Libraries* that article did not meet the requirements for the assignment (it is only three pages long). Because of this limitation, I browsed the same journal issue and located this article. I searched for the journal title in e-journals on the Hagerty Library website and navigated to ProQuest. I browsed the issue containing the Baru article and located the Zimmerman article.

Database: ArticleFirst [FirstSearch].

Method of Searching: Browsing [executed 3/3/2012]

Search String: Browsed the journal in which the following non-reviewed article was found:


Scholarly/Refereed Status: Ulrich’s indexes this journal as refereed, describing its content as academic/scholarly. The journal website gives instructions for reviewers, indicating that the articles in this journal are peer-reviewed. LISA does not index this journal. Taking these facts into consideration, I determined that this article is peer-reviewed and classified as a scholarly work.

Concluding Personal Statement

I feel that I learned a lot from this project. It is amazing how huge the amount of literature there is on this subject. I only scratched the surface with my initial searches. There were numerous citations that I could have followed if there had been more time for the project. It seems that I only scratched the surface of this topic. I enjoyed researching this topic – so much that I had to yank myself back to keep to the assignment. Though there was quite a bit of information on the curating aspect of e-science, there was an even vaster amount of information on the technical infrastructure necessary for e-science. Working on this project has made me more interested in metadata and the contribution to scholarship that good metadata can give. It also made me realize the importance of saving the raw data from science projects. I consider data curation important for the advancement of science and the good of society.
Absolutely outstanding job on your project. I was impressed with the variety of searching methods you used, ensuring that you only included articles that met our required criteria. You found excellent articles that focused exactly on what was required. Way to go! You have caught on well to this searching business! You earned 35 out of 35 points!

I certify that:

- This assignment is entirely my own work.
- I have not quoted the words of any other person from a printed source or website without indicating what has been quoted and providing an appropriate citation.
- I have not submitted this assignment to satisfy the requirements of any other course.

Signature  

Date  

Helen Helm  

March 16, 2012