Maria E. Chastain. Scientific Digital Data Repositories: Needs and Challenges for Cancer Researchers. A Master's Paper for the M.S. in L.S degree. June, 2013. 59 pages. Advisor: Helen R. Tibbo

The purpose of this study is to understand the varied data needs of molecular level cancer researchers who use light, fluorescent, and electron microscopy to obtain knowledge about cancer on a molecular level. It explores what data tools a sample of researchers are currently using to preserve their data for future access, and the needs of these researchers for depositing their digital research data into digital repositories. Data from the researchers suggest that they understand the need to preserve their raw and compiled data in places outside their laboratory, but they have not fully embraced the idea of depositing it in a repository. This seems most likely due to them not fully understanding what repositories are and what they provide. To increase the use of repositories by this research community, repositories need to promote themselves better and to offer additional services that are specific for the needs of this community.

Headings:

Digital preservation

Institutional repositories

SCIENTIFIC DIGITAL DATA REPOSITORIES: NEEDS AND CHALLENGES FOR CANCER RESEARCHERS

by
Maria Eugenia Chastain

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information/Library Science.

Chapel Hill, North Carolina

June 2013

Approved by

Advisor: Helen R. Tibbo

Table of Contents

Introduction

It is estimated that over 500,000 men and women will die from cancer in 2013 and another 5,000,000 will be told that they have cancer (American Cancer Society, 2013). To decrease the number of deaths due to cancer (as well as reduced the incidence of cancer in the first place) numerous researchers have joined together to understand every aspect of cancer from how it ravishes the body to how to use various treatment regimens to try to eliminate the cancer from a patient's body. The data that these researchers produce can be in the form of radiological images (projection, fluorescent, computed tomography, ultrasound, magnetic resonance, nuclear medicine), photos of the cancer in the body or when it was biopsied, histopathological sections, or molecular information about the cancer (what proteins is it expressing and their RNA transcript level, whether the gene encoding the protein has any mutations associated with it, etc).

Each type of data (radiological, histopathological, or molecular) has to be processed and analyzed by software that is often unique for that type of data and the data needs to be stored in such a way that future researchers can have access to this data. Research has focused on the best way to preserve radiological and histopathological data and what metadata is needed for allow future researchers to understand how that data was generated and processed. However, little has been reported on the varied data management needs of molecular level cancer researchers who work with a wide array of methodologies, technologies, and data types. Research is also lacking as to what is needed for that data to be preserved and accessed in the near and distant future. The latter is critical for researchers who acquire funding from the National Science Foundation (NSF) as well as from the National Institute of Health (NIH) as these funding agencies require researchers to have a protocol in place to preserve their research data for future reuse.

Because few scientists are trained to be digital data curators or managers, they do not have the skills to curate their own data. Moreover, they often lack the resources necessary for building robust information management and preservation infrastructures or to hire professionals for data management and data curation. Digital repositories, such as campus institutional repositories can provide a place to deposit data and to have them professionally maintained and curated. Many questions, however, persist as to researchers' data management needs prior to having their data ingested into repositories.

The purpose of this study is to understand the varied data needs of molecular level cancer researchers who use light, fluorescent, and electron microscopy to obtain knowledge about cancer on a molecular level. It explores what data tools a sample of researchers are currently using to preserve their data for future access, and the needs of these researchers for depositing their digital research data into digital repositories. Goals of current study are to 1) help target researchers understand the issues around curating their data, and 2) increase the likelihood that digital data generated by these researchers maintains its integrity over time.

Literature review

The Need for Scientific Research Data Management and Curation

A wide range of needs are driving the recent interest in research data management in the higher education, government, and commercial sectors. This paper focuses on work done within academic settings. Griffiths notes that a there are huge time and monetary investments by universities and other research organizations into research activities:

> "Across the spectrum of subjects and disciplines, researchers create and collect many different kinds of data during the course of their research. Data sets are generated through different processes and methodologies, for different purposes and beneficiaries" (Griffiths, 2009, p. 48).

Beagrie explains that the cost of data produced "through sensors, experiments, digitization and computer simulation" is very high noting in particular that "Satellites, particle accelerators, genome sequencing, and large-scale digitization and electronic publishing collectively represent a cumulative investment of billions [of dollars] in digital research and learning" (Beagrie, 2008, p. 7).

Besides the high cost of producing research data, another significant problem is the rapid growth of research data volume and its complexity (Gershon, 2002; Hey and Trefethen, 2003; Jirotka, et al., 2006; Borgman, et al., 2007; Kowalczyk, 2011).

In the last ten years there has been a marked increase in the number of institutions that are interested in creating an information cyberinfrastructure for data sharing, preservation, re-use, re-analysis, and data mining of the research data. Research data has significant scientific value and can be reused "to fuel new ideas and insights" (Kowalczyk, 2011, p. 1). Kowalczyk notes that "research data is an integral part of the scientific record as evidence of the rhetorical structure of scholarly communication", and availability of the research data is "necessary for replication and validation of scientific results" (Kowalczyk, 2011, pp. 1-2). Publishing research data affords researchers opportunities for collaboration (Heery and Anderson, 2005, p. 20; To share or not to Share, 2008, p. 26); greater visibility for their research groups and institutions; esteem factors and positive feedback to the local, state, and federal agencies, donors, and private organizations that funded the research (To share or not to Share, 2008, p. 26; Griffiths, 2009).

What Is Digital Research Data Curation?

Many scientists recognize the problem that the data within data sets need to be properly cared for to ensure that they retain their authenticity and understandability in order to remain useful resources. In light of these needs, the practice of digital data curation is emerging.

Digital data curation is a term that is used "for the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users" (Beagrie, 2008, p. 4)[1].

"Digital Curation" was first used publicly in London on the 19th October 2001 at the "Digital Curation: digital archives, libraries and e-science seminar" sponsored by the Digital Preservation Coalition and the British National Space Centre. The purpose of the seminar was to promote a "cross-sectoral dialogue between archivists, library and information management specialists, and data managers in e-science" (Beagrie, 2008, p. 4). One type of Data Curation, which is called Digital Research Data Curation, is

[1] **Beagrie** cited **Giaretta, D. (2005)**. *DCC approach to digital curation, version 1.23, 2005*, May 28; and **Joint Information Systems Committee. (2003)**. *JISC Circular 6/03 (revised): An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications.*

especially of interest to many researchers of different areas as it is “a multi-faceted issue, requiring technologies, organizational structures, and human knowledge and skills to come together in complementary ways” to ensure that research data maintains its provenance, integrity, and accessibility (Mayernik, 2012, p.1).

“Big Science” vs. “Small Science”

Lately, a lot of attention has been devoted not only to curation of data from “big science” that is generated by large research groups and often an array of sensors, but also to data curation from “small science” that is generated from small research groups whose data management relies upon the skills and knowledge of the individual researchers within the group (Martinez-Uribe, 2007; Kowalczyk, 2011, p.19; Normore and Tebo, 2011). Normore and Tebo (2011) defined “big science” and “small science” the following way:

> “Big science fields, such as physics and astronomy; are most often associated with large scale projects, large quantities of data often gathered from automated, sensor-derived sources, and impressive funding. “Little science” / “Small science” projects ... are more often associated with small research groups”.

Small science data is “less likely to be preserved in the long run”, “less likely to exist in standardized formats that facilitate immediate reuse”, “more likely to be heterogeneous in format and in need of “individual curation” than big science data”, and “metadata creation tools are not as common within the small science realm” (Dietrich, 2010, p. 81). In addition, some of the data generated by these groups are highly sensitive, and strict ethical and security protocols are needed to collect these data (Martinez-Uribe, 2007).

Preserving Research Data: The Records Continuum Approach vs. the Information Lifecycle Approach

There are methodological arguments about digital research data curation among supporters of the records continuum approach (McKemmish, 1997; Flynn, 2001; Bantin, 1998; Wilson, 2010) and the information lifecycle approach (Hodge, 2010). The main difference between these two approaches is in the recordkeeping and archiving processes. According to the lifecycle model there is a clear distinction between them.

> "The lifecycle model sees records passing through stages until they eventually "die", except for the "chosen ones" that are reincarnated as archives" (McKemmish, 1997).

The records continuum model sees recordkeeping and archiving as an integrated process:

> "The basic idea is that records can function both actively in the organization in which they were created and passively as part of an archive. There is no question of transferring or capturing records from an active environment into a passive archive" (Doorn and Tjalsma, 2007, p.8).

A central premise of the records continuum model is that "electronic archivists should be involved in the early design stages of new information systems" (McKemmish, 1999) and is very popular among researchers who specialize in digital research data curation (Lynch, 2008; Cragin, 2010; Kowalczyk, 2011). At the same time some researchers have pointed out that the records continuum approach was created for bureaucratic records which are completely different from the "often chaotic or anarchistic research environment", and implementation of the records continuum approach into scientific research practice is extremely difficult or impossible (Doorn and Tjalsma, 2007, p.9). The main criticism of the data lifecycle approach is that it pushes the responsibility to preserve digital research data to archivists in the future. Because the archivists will not have worked with the original researchers and because the data may not be well-prepared the preservation process difficult. Doorn and Tjalsma capture the dilemma well. On one hand data creators are seldom the ones responsible for long-term preservation of the data thus

> "they are not (or do not feel) responsible for the data after the research project for which they were created has finished. On the other hand, the organizations that are set up to take care of the long-term preservation (e.g. data archives) have hardly any influence over the creation of the data" (Doorn and Tjalsma, 2007, p.9).

Digital Research Data Curation of Scientific Data as an Object for Scientific Study

Due to the mandates and wishes of the funding agencies and journals that fund their research and publish their articles, respectively, scientists today are not only responsible for producing data, analyzing the data and making scientific reports about the results, but are also responsible for data management and curation, including data sharing and creation of the contextual metadata (Kowalczyk, 2011, p. 18). To help scientists with this task, some funding agencies, journals, university libraries (as well as data centers, and other Information Institutions) are "developing tools and services that enable researchers to manage, preserve, find, access, and use data within and across institutions and disciplines" (Mayernik, 2012, p.3). While a step in the right direction, these tools for research data management are still "expensive in terms of both personnel and equipment" for small science researchers and these scientists are forced to look for "necessary funding which would allow them to develop a robust data management infrastructure" (ARL, 2006; cited by Kowalczyk, 2011, p.24) or to use funds that are critically needed for their research.

In addition to the tools mentioned above, organizations with mission to help researchers with data management are also creating tools and theoretical models. Among them the most noteworthy are the Digital Curation Centre (DCC) (UK), the Virtual Research Environment (VRE) (UK), the Semantic Web and Autonomic Computing Programme (UK), and The Cancer Imaging Program (CIP) (USA). The Digital Curation Centre (DCC) concentrates on studying issues concerning scientific digital data curation based on the life cycle approach, "testing and evaluating tools, methods, standards and policies in realistic settings and offering a repository of tools and technical information" (Heery and Anderson, 2005, p. 7).

The Virtual Research Environment (VRE) programme "includes projects that will investigate building services on repositories to support research activity" (Heery and Anderson, 2005, p. 8). "The project utilised and evaluated the application of the records continuum theory to the practical management of digital records and their associated systems" (Heery and Anderson, 2005, pp. 8-9).

The Semantic Web and Autonomic Computing programme is "investigating ways to link e-prints and peer-reviewed articles to the primary research data upon which they are based" (Heery, et al., 2004, p. 8). For their proof-of concept, they linked primary data relating to crystal structures with e-prints which discussed the crystals via a "Resource Discovery Network science portal (PSIgate)". The project also shows the benefits associated with placing research data in an open access institutional repositories. One benefit is "making the data available for sharing and reuse in a timely fashion without the delay inherent in linking research data dissemination to the traditional journal publishing process" (Heery, et al., 2004, p.9).

The Cancer Imaging Program (CIP) branch of the National Cancer Institute (NCI) is trying to facilitate clinical decision making by the establishment of the Quantitative Imaging Network (QIN) (Levy, et al, 2012, p.1250). This network is to support the endeavors of the research institutions currently funded by CIP/NCI as well as the development of a research network that promote the sharing expertise, data, and technologies. "One of the goals of data sharing in this context is to enable secondary reuse of research data for validation of imaging algorithms and qualification of quantitative imaging biomarkers" (Levy, et al, 2012, p.1250).

Digital Repositories

There are different classifications of digital repositories in the literature. Heery and Anderson (2005, p.13-14) separated repositories into four different groups by content type, by coverage, by primary functionality of the repository, and by target user group. Content types include raw research data; derived research data; full text pre-print scholarly papers; full text peer-reviewed final drafts of journal/conference proceedings papers; e-theses; full text original publications (institutional or departmental technical reports); learning objects; corporate records (staff and student records, licenses, etc). Coverage, or collecting breadth of repositories, may be focused on personal (author's personal archive); journal (output of a single journal or group of journals); departmental; institutional; inter-institutional (regional); national; international. Repositories may also be classified by primary functionality of repository: (enhanced access to resources (resource discovery and location); subject access to resources (resource discovery and location); preservation of digital resources; new modes of dissemination (new modes of publication); institutional asset management; sharing and reuse of resources. Finally, one can group repositories based on audience of the target user groups: learners, teachers, researchers. Kowalczyk (2011, p.151) united repositories into three groups: research (institutional, work unit, commercial storage, personal); community (domain repositories, journal repositories); and reference (national repositories, journal repositories). Nicholas *et al.* suggested the following classifications: institutional repositories; subject repositories ("based on collecting only within a certain discipline"); and, format repositories ("scope is limited by collecting in a particular format, e.g. student dissertations and e-theses, research data, digital images") (2012, p. 196).

In 2003, Lynch defined a university-based institutional repository as "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members." Among services he suggested it should provide are long-term preservation, organization, access, and distribution. In addition he remarked that "a key part of the services that comprise an institutional repository is the management of technological changes, and the migration of digital content from one set of technologies to the next" (Lynch, 2003, p. 328). Lynch's definition of institutional repositories as a set of services influenced many researches as to what an institutional repository should be. Extending from Lynch, Cragin and Shreeves defined an institutional repository as "a set of services and technologies that provide the means to collect, manage, provide access to, disseminate, and preserve digital materials produced at an institution" (2008, p. 89).

Under the "Institution" umbrella Cragin and Shreeves envisioned a wide spectrum of organizations: colleges, universities, governmental agencies, museums, corporations, and other organizations (2008, p. 89). Heery and Anderson concentrated their attention on aggregation of metadata exposed by institutional repositories, and described the following services that are provided by institutional or subject (‘themed’) repositories: indexing the content of repositories; impact analysis and provision of other metrics with regard to content; metadata enhancement services; metadata creation; annotation services; supporting accountability for the “evidence base” of data produced in laboratories; recording health and safety information (2005, p. 4-5). Martinez-Uribe spoke about repository services not as a technical infrastructure “that allows storage, access, description, dissemination and preservation of digital objects”, but also as a support in resolving legal issues, and development policies for “the creation, deposition and sharing of digital research outputs” (2007, p.30). Kowalczyk believed a repository should provide the following services: backups, contingency, planning, process resumption planning, hardware and network redundancy, automatic failover, and site mirroring (2011, p.21).

Barriers to Depositing Research Data into Repositories

For many scientists, data management and deposit of data into a repository for reuse are not key issues. While many researchers see the value in sharing data many do not want to take the time necessary to prepare the data for long-term preservation and others simply do not want to share their data with the public or with other researchers.

Barriers to depositing research data into repositories can be divided into 1) personal or subjective barriers and 2) objective barriers. One of the subjective barriers concerns the ownership of the data within a repository, especially if the data was a result of a collaborative research project between many institutions (Martinez-Uribe, 2007). Additional subjective barriers include the fact that data management is time-consuming and costly; it take time and effort to document data collection, workflows and analyses; robust metadata is needed for data reuse; data management is unfamiliar to most researchers; researchers are often uncertain as to where to archive data; some researchers fear professional competition and being “scooped” in their research; and some researchers fear others will exploit their hard work when reusing their data. (Schröder, 2007, p.68; Griffiths, 2009; Kowalczyk, 2011, p.173; Langer, 2011, p.203). There is also a fear of incorrect or inappropriate interpretation of the deposited data (Cragin, 2010), and the misuse of the data, such as the selective use some of the data points while not using other data points (Borgman, 2012, p.1059). Some objective reasons as to why research data cannot be deposited are that the size of files may be too large for file-sharing service (Cragin, 2010), or data cannot be shared by ethical or epistemological reasons (Borgman, 2012, p.1060).

Data Sharing

There are also arguments in literature about the effects of data sharing on science. Advocates of data sharing frequently point to the possibility to reproduce and verify research, thus enabling “others to ask new questions of extant data”, “to advance the state of research and innovation” (Borgman, 2012, p.1059), to improve the quality of data, and to increase the usage of standards (Karasti, et al., 2006, p.348; Schröder, 2007, p.68). Opponents argue that the process of reproducing results depends upon many factors, which may or may not be not known to the researcher who tries to reuse/or reproduce the data. Some of these factors relate to details about the instrument used to gather the data (such as how it was calibrated, lab-specific practices, and the settings and parameters associated with the commercial software needed to run the instrument) (Borgman, 2012, p.1060) while other factors relate to specific methods and procedures used when collecting the data (Borgman, 2012, p.1070). Opponents also say that if one uses shared data, they must be cautious about the quality and reliability of the data they are using. Some point to the research conducted by Ashelford *et al.* as a prime reason for this cautiousness. Ashelford *et al.* after re-analyzing 16S rRNA sequence records that were deposited in public repositories found out that at least 1 in 20 records contained "substantial anomalies" (Ashelford, *et al.*, 2005, p. 7724).

To counteract some of this negativity and to alleviate the concerns about sharing research data with the public, the Research Information Network (RIN) group has tried to encourage data publishing and their reuse by the active promotion of case studies, career-related and grants funding rewards, information support, and the availability of expert advisers (To share or not to Share, 2008, p. 9-10).

In addition, Parsons *et al* predict that under pressure of funding agencies the number of scientists who will deposit their research data will significantly grow. They envisage that one day scientific data will be "highly distributed and housed at many different types of institutions [and that] … the use and users of the data will be very diverse and even unpredictable" (Parsons, et al, 2011, p. 565). Moreover, they point out that while they foresee that data will become more available, the different data formats and vocabularies of the data "will continue to be very complex if not chaotic" (Parsons, et al, 2011, p. 565).

There has also been discussion about formal versus informal ways of sharing data. Heery and Anderson have postulated that data sharing based on informal methods (such as sharing data through owner created and managed websites, informal networks, wikis, and peer-to-peer mechanisms) in near future may prevail over formal ways of sharing data via digital repositories (2005, p. 20)

Due to the fact that a number of researchers are not comfortable with using data repositories for their own data or sharing their data with the public, data repositories have been under-utilized. The literature includes several suggestions as to how to improve data repository usage and increase data sharing and data reuse.

- Make digital repositories researcher-contributor friendly (Heery, and Anderson, 2005, p. 4).
- Extend the number of services provided by the repositories which are flexible and controlled by scientists, such as an embargo service (Cragin, 2010, p. 4035-4036), social communication / collaborative environment (Cragin, 2010, p. 4035-4036; Osswald, 2008, p. 520), data management consultations (Cragin, 2010, p. 4035-4036; Martinez-Uribe, 2007), format and content migration, metadata support (Riding the wave, 2010, p.20), indexing, automated subject classification, and name authority services (Heery and Anderson, 2005, p. 5; Osswald, 2008, p. 519).
- Establish collaborative relationships with other repositories (Cragin and Shreeves, 2008, p. 93).
- Develop data citation and persistent identification capabilities (Osswald, 2008, p. 519; Riding the wave, 2010, p.20).
- Create the possibility to store a large volume of data along with the ability to restrict who can share, delete, and edit the information (Martinez-Uribe, 2007; Riding the wave, 2010, p.20).

- Develop ethical frames for data creators and data users (Parsons, et al, 2011, p. 566).

Methodology

An email survey was conducted to understand the needs of cancer researchers who study cancer at a molecular level in regards to depositing their digital research data into digital repositories. The survey was constructed from open-ended, close-ended, polar and contingency questions. The following operational definitions were used during the construction of the survey: file format, data set, research data quality control criteria, contextual metadata, and digital repository.

Terminology Used

● *File format* was defined as the structure and type of information stored in a file. The structure of a typical file may include a header, metadata, saved content, and an end-of-file (EOF) marker. File formats may either be proprietary (for instance the .doc format) or open (such as the .txt format).

●*Data set.* There is no single well-defined concept of what a dataset is. Renear *et al.* note that "the variations in individual terms are significant, the terms themselves are often used in different senses, and critical characteristics are left underdetermined" (Renear, *et al.*, 2010, p.3). For the purpose of this study, data set was defined as a collection of related data records on a storage device.

● *Research data quality control criteria* was defined as a set of rules or principles used for evaluating, testing, or ensuring the conformance of data values - some of which are consistency, completeness, accuracy, and precision.

● *Contextual metadata* - was defined as "the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record" (Grace et al., 2009, p. 4). The process of creation of contextual metadata is based on the information lifecycle approach (Lee, 2011, p. 117, Ixchel and Yakel, 2011) and included nine classes of contextual entities object, agent, occurrence, purpose, time, place, form of expression, concept/abstraction, and relationship (Lee, 2011, p. 95).

● *Digital repository* was defined as a set of services and technologies that are provided by an institution. Examples of organizations that were defined as an "institution" are colleges, universities, governmental agencies, museums, and corporations (Cragin and Shreeves, 2008, p. 89-90). Examples of services and technology that could be provided by an institution are storage, access, dissemination and preservation of digital objects, development policies for "the creation, deposition and sharing of digital research outputs" (Martinez-Uribe, 2007, p.30), backups, contingency, planning, process resumption planning, hardware and network redundancy, automatic failover, and site mirroring (Kowalczyk, 2011, p.21), description, indexing the content of repositories; impact analysis and provision of other metrics with regard to content; metadata enhancement services; metadata creation; annotation services; supporting accountability for the "evidence base" of data produced in laboratories; and, recording health and safety information (Heery and Anderson, 2005, p. 4-5).

Email Survey

The researcher elected to conduct an email survey due to its many strengths but was also aware of the methodology's weaknesses and limitations. The strength of using email surveys are the following: 1) there is time flexibility for the participants to answer the questions, 2) the location of the participants does not matter, 3) they are less expensive than face-to face or telephone interview instruments, 4) they enable participants to provide “well-thought-out answer rather than top-of-the head responses” (Hart, 2006, 5), they allow the researcher to interview more than one participant at a time, and 6) they eliminate the need and cost associated with transcribing and editing the answers since they were already in electronic form and required little editing or formatting before being processed for analysis (Meho, 2006, p. 1288). Weaknesses include potentially low response rates especially with open-ended questions and overly brief responses to open-ended questions.

Population

The population of this study was fifteen cancer researchers from five institutions (University of Cincinnati, Indiana University at Bloomington, University of North Carolina at Chapel Hill, Georgia Tech, and Emory) who use light, fluorescent, and electron microscopy for generating their research data. They all have been or currently are Principal Investigators (PIs) and have had two or more people in their research lab. They were selected based on their specific research area and were identified through their research.

Data Collection Procedures

Potential participants were sent an introductory email that contained a consent to participate in a research study form (i.e., a consent form). This consent form had three sections. The first section contained the IRB study number, title of the study, the name of graduate student who was conducting this research for her master’s paper, contact information of graduate student, the name of the faculty advisor and her contact information. The second section stated that participation in the study is voluntary and without monetary compensation. The third section described the study, the purpose of the research, the population associated with the survey (i.e., who was being interviewed), the number of people who were part of this research study, the length of time one had to fill out the survey (i.e., the duration that the study was going to be open), and the potential benefits and risks associated with their being part of the study. At the end of the consent letter was an area for the participant's name, signature, and date that they signed the form. The email text had the same information as the attachment as well as a note thanking them for their time. Cancer researchers who agreed to participate in the study and who signed their consent forms were sent the questionnaire. One week after the questionnaire was sent to them, a follow-up email was sent to those who did not send back their survey to politely remind them to complete and send back the survey.

Data analysis procedure

After the survey was finished and sent back to the investigator, the survey data was analyzed. Because the questions are predominantly open-ended, content analysis was used to make sense of the participants' responses. The assessments that were made are the following:

1. Are there any commonalities across respondents to these questions?

2. Are there any remarked differences across respondents to the survey questions?

3. What is the percentage of respondents who will use a repository a) under the current setup, b) with tweaks with how the repository does things, c) under any conditions?

4. What are the common concerns about the usage of repositories?

Data Handling Procedure

The participants were not asked for any personal information, but because the survey included open-ended questions the respondents might have include self-identifying information in their answers. To ensure confidentiality the responses were separate from any personal information and any self-identifying content was removed from responses. Attachments containing questionnaires and their answers were immediately saved on a personal computer and the emails which contained the attachments were promptly deleted as well as purged.

Limitations of the Study

The main goal of this study was to provide valuable insights about needs of a unique subset of cancer researchers, those who study cancer on a molecular level using light, fluorescent, and electron microscopy, regarding depositing their digital research data into digital repositories. Due to the narrow scope of those who were being surveyed, the study is not appropriate for deriving statistical descriptions for all cancer researchers nor is it appropriate to generalize across the entire population of cancer researchers.

Results

The scientific digital data repositories: needs and challenges for cancer researchers survey was administrated between May 20, 2013 and June 7, 2013. The participants of this study were fifteen cancer researchers from five institutions (University of Cincinnati, Indiana University at Bloomington, University of North Carolina at Chapel Hill, Georgia Tech, and Emory University) who have been or currently are Principal Investigators (PIs) and who have had two or more people in their research group. All chosen PIs from the five institutions agreed to participate in the current study and returned a completed consent form and the survey (See Appendix I. Survey and Appendix II. Consent Form).

The first two open-ended questions of the survey were designed to gather two demographic factors: the scientific domain (area/focus and the methodology they use) and the size of laboratory (See Appendix I. Survey). Content analysis of the answers to the first question revealed that some of them worked in more than one research area. Most of the PIs (seven) work within the Molecular Biology domain, four PIs indicated Genetics, four PIs - Cell Biology, two PIs – Biochemistry, two PIs – Microbiology, one PI – Biophysics, and one PI – Cancer Biology (see Table 1). All of the respondents focus on different areas of cancer research: imaging in cell biology and biophysics, DNA visualization, DNA repair, genome stability, chromosomal instability in cancer, DNA replication stress, telomere structure, telomere replication, DNA and RNA isolation, qPCR, methylation analysis, microRNA determinations, transfection of cells in culture, mechanisms of mutagenesis and carcinogenesis, and eukaryotic cell cycle. Three groups of PIs concentrate on such topics as DNA damage, DNA fibers, and DNA sequencing.

The number of people within a lab varies from 0 to 14. Two PIs have three researchers in the lab, and five PIs have five researchers in the lab (see Table 2).

Table 1. PIs' Scientific Domain

Domain	Responses
Molecular Biology	7
Genetics	4
Cell Biology	4
Biochemistry	2
Microbiology	2
Biophysics	1
Cancer Biology	1

Table 2. Number of Researchers in the Lab

Lab Size	Responses
0	1
1	1
2	1
3	2
4	1
5	5
6	1
7	1
8	1
14	1

All PIs are doing their research within different scientific domains and focus upon different aspects of cancer. They also have different numbers of researchers in their labs. The difference in number mainly seems to be dependent upon lab funding. The PIs responses reflect the current status of their lab activities. The lab personal distribution does not show the maximum number of people in the lab during well-funded periods.

The third open-ended question was about file formats which PIs use in their researching work on a regular basis. All file formats described in the returned surveys are categorized into the following groups:

- Presentation (.ppt, .pptx, .key);
- Collaboration (.wsp);
- Multimedia format (.mov);
- Image file formats (.jpg, .tif, .bmp, .gif, .png, .pct, .czi, .lsm, .gel).
- Graphic file formats (.eps, .ai)
- File formats that can be saved as a plain text format (e.g.: .doc, .docx, .xls, .xlsx, .ics, .ab, .gb, .fcs).

The most popular file formats were Word document (.dox/.docx, Microsoft Office) (15 responses), Tagged Image File Format (.tif, Adobe) (15 responses), Image file format (.jpg, Joint Photographic Experts Group) (12 responses), and Excel spreadsheet (.xls/.xlsx, Microsoft Office) (11 responses) (See Table 3). The analysis of file formats mentioned by the PIs showed that the majority of the formats are produced by commercial software firms such as Microsoft, Adobe, Zeiss, Gel Doc and Apple, (see Table 4).

Table 3. File Formats

#	File format extension	Responses
1	.doc / .docx	15
2	.xls / .xlsx	11
3	.ppt /.pptx	5
4	.key	1
5	.wsp	1
6	.jpg	12
7	.tif	15
8	.bmp	1
9	.gif	1
10	.png	1

#	File format extension	Responses
11	.pct	1
12	.mov	1
13	.pdf	7
14	.psd	2
15	.eps	1
16	.ai	1
17	.ics	1
18	java	1
19	.czi	1
20	.lsm	3
21	.gel	3
22	.ab	1
23	.gb	1
24	.fcs	1

Table 4. Description of the File Format Extensions

#	File format extension	Description
1	.doc / .docx	Word document (Microsoft Office)
2	.xls / .xlsx	Excel spreadsheet (Microsoft Office)
3	.ppt /.pptx	Power Point presentation (Microsoft Office)
4	.key	Keynote presentation (Apple)
5	.wsp	Windows Sharepoint (Windows)
6	.jpg	Image file format (Joint Photographic Experts Group)
7	.tif	Tagged Image File Format (Adobe)
8	.bmp	Device independent bitmap file format (free of patents)
9	.gif	Graphics Interchange Format (patents are expired)
10	.png	Portable Network Graphics (free of patents)
11	.pct	Picture Image file (Apple)
12	.mov	Quick Time File Format (Apple)
13	.pdf	Portable Document Format (Adobe)
14	.psd	Photoshop Document file (Adobe)
15	.eps	Encapsulated PostScript (Adobe)
16	.ai	Adobe Illustrator vector graphic file (Adobe)
17	.ics	iCalendar file format
18	.czi	Microscope image data file (Zeiss)
19	.lsm	Image file format for LSM (a line of laser point scanning confocal and two photon microscopes) (Zeiss)
20	.gel	Image file format for Gel Doc EZ system (automated gel imaging instrument)
21	.ab	DNA Sequence file
22	.gb	GenBank sequence database file
23	.fcs	Flow Cytometry Standard file format (Standards Task

#	File format extension	Description
		Force (DSTF))

The fourth question was closed-ended and concerning the type of software (commercial or lab derived) that the PIs use for generating their research data. All PIs who responded to this question use commercial software (15 responses), which can be categorized into the following groups: microscope imaging software; image processing software; data generation software; and, software which accompany the laboratory equipment. Two PIs use also lab derived software for image processing and image analysis and another two PIs use open source software.

The fifth question was a closed-ended question and about the type of software (commercial or lab derived) in which the PIs use for analyzing data. Most scientists use commercial software such as:

- Microscope analysis software;
- Statistical analysis packages;
- Graphing software; and
- Software for analysis of sequencing data and for genetic engineering projects

Some PIs combine commercial with open source software during the process of analyzing data (3 responses). One PI uses only open source software and four PIs use lab derived software.

The sixth question was an open-ended question about the average size of the data sets that PIs generate during an experiment/study. Every PI described datasets differently. Seven PIs who responded gave an average, total size of their datasets. One PI defined average size of dataset in the KB range; four PIs in the MB range, two of them recorded the particular size data sets between 10-20 MB; three PIs in the GB range (1GB (1 response), 2-3 GB (1 response), 10 GB (1 response)). Some PIs tried to give an average size for the different file formats (such as text files, DNA sequencing files) they use (2 responses).

Seven PIs gave a range to their data size. The reason for this variability is that each experiment they do varies in complexity, scope, and scale.

- from KB to MB (1 response);
- from KB to GB (3 responses);
- from KB up to 2TB (2 responses);
- from MB to GB (1 responses).

The seventh question consisted of two parts. The first part was a polar question about data quality control criteria (needing a “yes’ or “no” answer): “Do you have a set of formalized quality control criteria to ensure that the integrity of the data and files that you have remain intact and accessible?” The second part was a contingency question asked them to expand upon their initial answer, if it was a “yes”. Seven PIs gave short “no” answers; six PIs described how they save and backup data. For this purpose three PIs use a dedicated lab server, two PIs use different hard drives, and one PI uses a departmental server. One PI in addition to a lab server also uses a commercial clouding service (Amazon) for backup data. One PI has an “informatics group” which is responsible for data archiving and providing access to data to all lab members. One PI pointed out the formal barriers for alteration of data such as system restrictions for viewing, adding, editing, and deleting data. Only one PI described a set of formalized quality control criteria, which include biological controls, replicate controls, and machine controls (when applicable), and also “a few analytical measures that are employed to ensure that chemical process steps were complete (for methylation analyses)”.

The next set of questions (8 – 15) was about contextual metadata. These question ask whether the PIs describe experiments and if so where was the description, whether they have special lab templates for any descriptions that they have, how many descriptive elements they usually use for description of individual data set and data collections, do they use some metadata standards, if they generate metadata automatically, about the importance of explanation how the research data was generated and what kind of methodologies were used for it.

The eighth question in the survey consisted of two parts. The first part was a polar question: “Do you describe your experiments, the conditions used during the experiments, the parameters used during data acquisition, and the changes that occur when you process the data”? The second part was a contingency question and was designed for PIs who responded “yes” to the initial question: “If so, is this description in electronic or paper form”? Only one PI answered the first part of the question “no”, and one PI answered “not sure”, with the remaining thirteen PIs indicating “yes”. Among those thirteen PIs, three of them use electronic forms of experiment description, six use the paper form (mostly lab notebooks), two use both electronic and paper forms, and two indicated that they describe experiments in published works.

The ninth polar question was an extension of the eighth question and was about whether the PIs had special lab templates for the description of research protocols and analysis. Of the fifteen PIs whom answered this question only four answered “yes”, while the remainder answered “no”.

The tenth and eleventh questions were open-ended and concerned the quantity of metadata elements describing individual data sets and data collections. The results for both questions were similar (See Table 5). Two PIs did not answer or typed “0” for both the individual data set and data collections questions. Three PIs indicated the number of metadata elements for an individual data set they use is between one and three while four PIs used the same number of metadata elements for data collections. Eight PIs used between four and ten elements for an individual data set while seven PIs used that number for data collections. One PI believes that the quantity of metadata elements has to be “as much as possible”. Another PI thinks that for both cases (an individual data set and for a data collection) the number of metadata elements should be “as many as necessary”.

Table 5. The Quantity of Metadata Elements for Description of Individual Data Sets and Data Collections

Quantity of Metadata Elements	Individual Data Sets	Sata Collections
0 / without answer	2	2
Between 1 and 3	3	4
Between 4 and 10	8	7
“as much as possible”	1	-
“as many as necessary”	1	1
“too vague”	-	1

The next three questions (questions 12 – 14) focused on metadata standards and were confusing for almost all of the PIs. The twelfth question was open-ended concerning what kind of metadata standards PIs used. All the PIs provided “negative” responses (see Table 6). The responses ranged from “none”, “not sure”, to “N/A”. Some PIs wrote: “No idea. Don’t know what this means”, “I do not understand the question”, “I don’t know what these are”. Responses demonstrated that all PIs were not familiar with the term “metadata”.

Table 6. Metadata standards

Question	Response "none" or "N/A" or "not sure"	Response "yes"
Which metadata standards do you use?	15	0

The thirteenth question consisted of two parts. The first part was a polar question (needed to be answered "yes" or "no"): "Do you generate metadata automatically?" Only two PIs answered "yes", the rest thirteen PIs answered "no", "not sure", "don't know", "I don't know what this is" (see Table 7). The second part of the thirteenth question was a contingency open-ended question asking PIs who answered "yes" on the thirst part of the question to explain the process of generation metadata automatically. From two PIs who answered "yes", only one pointed to the software that comes with confocal microscope.

Table 7. Automatically Generated Metadata

Question	Response "none" or "N/A" or "not sure"	Response "yes"
Do you generate metadata automatically? If so, what is the process?	13	2

The fourteenth question was an open-ended question about the format/file format in which metadata is preserved. There were four PIs who answered that metadata was preserved in the data file, and three researchers who noted that the metadata files were automatically generated by the confocal Zeiss microscope when the images were taken. The researchers were not asked to identify the brand of microscope that they use for generating their data, but most of them use either Olympus or Zeiss Both companies (Olympus and Zeiss) have software for automatic generation of metadata, but not all of the PIs mentioned this option.

Comparison of answers between one group of questions (questions 9 – 11) where the term "description elements" was used and another group of questions (questions 12 –

14) where the term “metadata” was used indicates that the majority of PIs do not associate descriptive elements and metadata with each other (even though they are essentially the same). Some potential reasons for the lack of connection between the two terms is that the PIs are not familiar with what metadata is, that metadata is associated with information and library science and not with their work, or perhaps they may think that their methodology, experiments, data sets, and collection of data sets are only necessary to help them understand their research and do not have to be describe for data reuse by other researchers. Most of the PIs create description of their data sets without using standardized forms, just to satisfy their own lab needs, whether these records will be able to facilitate future data reuse by other scientists remains an open question.

The fifteenth question was open-ended: “Who do you think should be responsible for metadata creation (e.g., researcher, data professional, librarian or someone else)?”, and elicited very strong opinion among the PIs. Nine PIs strongly stated that the researcher is responsible for metadata creation. One PI explained his/her point of view on this question the following way: “Until recently, I never thought about metadata. I knew that the images that we obtain had metadata that the microscope embedded. However, we have started to think about metadata and how to add it along each step of the processing of the images and subsequent analysis. With what I know now, I would say that ultimately it is a researcher’s responsibility”. Only two PIs thought that metadata creation is the responsibility of both researcher and data professional. One PI said it is a Data Professional’s responsibility. Three PIs answered “not sure” or “do not know”.

The sixteenth was a polar question about the PIs experience with depositing research data into a digital repository or data warehouse. This question highlighted the fact that the majority of them do not have any experience (11 responses) or that they had limited experience (3 responses). Only one PI had experience with a digital repository/data warehouse. The absence of experience may be explained by the absence of a developed information infrastructure for these researchers in their institutions or due to poorly promoted repositories.

The seventeenth question was a close-ended question concerning the way in which research data should be preserved and archived. Respondents were presented with the following choices: lab storage devices, institutional repositories, and community supported research collections. Six PIs had unambiguous answers: four PIs chose lab storage devices, one PI chose institutional repository, and one PI chose community supported research collections. Eight PIs were sure that research data should be preserved/archived in more than one location. Some of them chose all locations, lab storage devices, institutional repositories, community supported research collections (4 responses) and some of them on lab storage devices and in institutional repositories (4 responses). One PI suggested an alternative answer: to store everything in standardized formats on the cloud and provide open access to the data. The PIs' answers demonstrate their understanding of the importance of preservation and archiving their research data in multiple locations, not only on lab storage devices.

The eighteenth was a close-ended question concerning the type of data that should be preserved in research data digital repositories. The suggested variants were published, unpublished, raw data, and aggregated data files. Six PIs chose the variant “all data” and the remainder, nine PIs, chose “raw data”, six of those nine said that raw data from published results was what should be preserved. The answers provided by the PIs demonstrated that they understand the importance of preservation of different types of data, especially the raw data, but may be overlooking some privacy concerns unless the raw data is deidentified. Question 19 is an extension of this question and asked the PIs to explain why it is important for them to know how other researchers generate their data and what kind of methodologies they used.

The nineteenth question consisted of two parts: a polar question and a contingency open-ended question. In the first part PIs were asked whether or not publishing explanations as to how researchers generated data (instruments, experiments) and what methodologies researchers used was important. On this question all PIs answered “Yes”. In the second part of the question 19 the PIs were asked to provide details if they had replied to part 1 in the affirmative. Ten PIs’ indicated that experiments have to be reproducible and thus detailed methodology sections are essential, while two PIs’ wanted to be sure of the quality and validity of the data. One PI pointed out the importance of published explanations, “Because I … want to know how data was generated, processed, and analyzed so I don’t have to re-invent the wheel should I want to repeat those same experiments”. Five PIs’ discussed the role of peer reviewed journals in the detailed explanations of methods used by researchers. One PI stated that “[it] is extremely important and should be enforced by peer reviewed journals. All the parameters, instrumentations methodologies must be described meticulously”. However, another PI mentioned that the capability to write this type of detailed methods sections seems to be curtailed by the peer reviewed journals themselves as “journals have shortchanged the scientific community by minimizing methods sections and including them in the word count of the manuscript” and thus it seems that PIs are forced to refer readers of their paper, via citations, to other papers for a detailed explanation. In the eyes of one of PIs, these “simple references to the [other] literature are often insufficient”.

The twentieth question was a close-ended question that asked PIs about who should pay for the cost associated with storing the data/files in a repository. The respondents provided the following options: the researchers who publish the data, the researchers who access the digital repositories, commercial companies (e.g., pharmacies) who want to re-analyze research data, or should all data be completely free for everybody and supported by the local, state, or federal government. On this question, the PIs seemed equally divided at to who should pay for costs associated with storing the data/files in a repository. Five PIs said they were already paying for storage. Three PIs thought that users should pay for access to the data. Four PIs thought it should be free to non-commercial organizations. Nine PIs thought that the costs associated with storing the data should be defrayed in part or wholly by federal agencies, by funding agencies, or by the institutions (especially universities) in which the PIs work. One PI suggested the following option: “I think the cost should probably be paid for by the researcher who generates the data. But I also think there should be an allowance for it on grants. However, it would have to be set up so that it is mandatory and doesn’t come out of money that could be used for salaries or supplies… otherwise no one will be willing to pay. An affordable fee to use the data would be ok too but it would have to be affordable”. The PIs had quite different opinions as to who should pay the cost associated with storing the data/files in a repository. Being both researchers and lab managers they constantly have to deal with financial questions and budgeting, and thus see repository as a business entity that needs monetary support.

The question 21 consisted of two parts. The first part was an open-ended question about what kind of services digital repositories should have. The second part was a polar question asking if PIs are willing to use these services for a fee. The PIs believe that the data repositories should have the following services/capabilities: preservation; access to data files through metadata; access restrictions set by the data creator; search, inspect, download, find, related datasets; a research data management specialist(s) who will help them access data; cloud-based backups; and a way to easily enter, maintain, and access their data. On the second part of the question about a fee for a service, eight PIs answered “yes” and five PIs answered “No”/”Not sure”. As for who should pay for the service, seven PIs thought that the users should pay for the services, one thought that the user’s institution should pay. One PI seems pragmatic about who should pay for any service provided by a repository, “I think the cost should probably be paid for by the researcher who generates the data. But I also think there should be an allowance for it on grants. However, it would have to be set up so that it is mandatory and doesn’t come out of money that could be used for salaries or supplies… otherwise no one will be willing to pay. An affordable fee to use the data would be ok too but it would have to be affordable”.

The twenty-second question was closed-ended question and asked what kind of digital repositories should be inter-institutional, cross institutional, or national. Eleven PIs did not discuss what type of repositories should be those types of entities, but six of them believe that data within a repository, regardless of what type, should be freely available and open to everyone. One PI thinks that repositories should be inter-institutional. One PI saw this question as a complex problem: “Published data should be [deposited in a] national or international [repository]. International can be a problem… NCBI is good but why should the world have free access to something funded by the US government?" While another PI was a little more explicit in what data should go where: "Digital Repositories that house health data should be all three. Data Repositories which contain data generated with local, state, or federal funds should cross institutional and state. Data that a lab generates (but has not been analyzed or processed) should be inter-institutional".

Conclusion

Cancer Researchers are developing and using a variety of techniques and methods which employ light, fluorescent, and electron microscopy to understand the mechanism by which cancer occurs in a person, to detect a cancer in a person as early as possible, to determine the best treatment regimen that will eradicate a person's cancer, and to prevent the cancer from reoccurring (or a new one from forming). Very little is known about the types of digital data they generate, the metadata which describes how this data is generated, and how this data is preserved for future use/access. To shed light on these topics, 15 PIs were sent a twenty-two questions survey. They all completed survey and sent it back (a 100% response rate). The respondents are specialized in a number of disciplines, such as, molecular biology, genetics, cell biology, biochemistry, microbiology, and biophysics.

These researchers generate a range of data type and volume. The average size of their data sets varies from a few kilobytes (KB) to a few terabytes (TB). Most of the PIs use different methods to save their experimental and data acquisition parameters but less than 50% save these parameters in an electronic form. Regardless of the medium by which they save the parameters, a majority use between 4 and 10 elements to describe what they do to obtain the data as these elements are critical for them to reproduce/repeat the experiment. This finding is in agreement with Greenberg *et al.* who found that

> "Researchers prefer rich descriptive metadata supporting discovery and reuse, although they are not necessarily dedicated to allocating time required for creating good quality metadata" (Greenberg, et al, 2009, 196-197).

The rich descriptive elements associated with describing the parameters used in an experiment sometimes are not fully elaborated upon when the PIs publish their work in a peer reviewed journal nor is all of the data (particular the raw, unprocessed data). This seems to be due to the word and size constraints placed upon the PIs by the journals and not due to the PIs not wanting to be open with their data and research. Digital repositories seem to be an avenue by which PIs could fully disclose how and what they do in an experiment as well as a medium by which to share raw data or additional information.

One interesting finding that came from the analysis of the survey is that these PIs do not see the acquisition parameters as a form of metadata. One can think of a couple of ways to remove this misunderstanding, 1) PIs could seek out training about metadata, 2) students who are in training to be cancer researchers could take a course on metadata and the power of its use, and/or 3) metadata specialists could give a series of lectures (or at least one) about what metadata is and how they can use it. The lack of familiarity with metadata coupled with the continued reliance on lab notebooks may be one of the major reasons why these PIs, and perhaps cancer researchers as a whole, do not use data repositories. Another reason why repositories are not highly used by the researchers maybe because the researchers believe that their institutions or granting agencies should pay for or help to pay for the cost associated with them using the repositories.

It seems that digital repositories could increase their use by these researchers, and most likely other investigators who do similar type of research, by doing a better job of advertising/promoting what they do and offer services that align with these PIs' needs. These needs include: easy access to their data; the capability to restrict who has access to parts of the data (namely access to data that has not been published); regular back-ups; and the ability to search, inspect, download, and find related documents. These needs are aligned with the primary functionality of a repository envisioned by Heery and Anderson (2005, p. 13-14) mentioned previously and thus do not seem excessive or out of the scope of what digital repositories are meant to do. It also seems that it would be beneficial for digital repositories to have a data curator available to assist researchers depositing their research data and retrieving other research data.

It is reasonable to expect funding agencies, which are forcing researchers to have their data available, to provide money to aid these researchers, and cancer researchers as a whole, in depositing their data into digital repositories. Defraying the costs for researchers to use a data repository seems especially important when some of the data elements collected can be terabytes in size which means it could cost a researcher $15,000 to $20,000 over 15 years just for one data element or >$100,000 for a data set (calculated according Plale, *et al.*, 2013). This cost would be prohibitive for most researchers.

In addition, digital repositories should expand their collaborative partnerships and start to work with owners of commercial software to create mechanisms to make available for review or re-analyzing the raw data. It is also important for repositories to establish relationships with commercial software vendors so that they can have copies of all (or at least a majority) of the software that their patrons used in order make it available for researchers when software will be no longer on the market.

Bibliography

Aiken, P., Gillenson, M.L., Zhang, X., and Rafner, D. (2011). Data Management and Data Administration: Assessing 25 Years of Practice. *Joumal of Database Management*, 22(3), 24-45.

AIM Pathology 1.0 / PAIS 2.0 Report. (2011). Center for Comprehensive Informatics, Emory University. Center for Biomedical Imaging and Informatics, the Cancer Institute of New Jersey.

Anderson, D., and Clark, S. (2006). Metadata, Contextual Data, and the Canadian Century Research Infrastructure. *The Serials Librarian*, 51(2), 117 - 126.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., Weightman, A.J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12), 7724 - 7736.

Bantin, P.C. (1998). Strategies for Managing Electronic Records: A New Archival Paradigm? An Affirmation of Our Archival Traditions? *The Archival Issues*, Vol. 23.

Beagrie, N. (2008). Digital Curation for Science, Digital Libraries, and Individuals. *The International Journal of Digital Curation*, Vol. 1, No. 1, pp. 3 - 16.

Biomedical Informatics Research Center in Stanford University. (2013). *Annotation and Image Markup (AIM) Project*. Retrieved from http://bmir.stanford.ed

Borgman, C.L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059 - 1078.

Bravo, B. R., and Díez, L. A. (2007). E-science and open access repositories in Spain. *OCLC Systems & Services*, Vol. 23, Iss: 4, 363 - 371.

Broeder, D., Kemps -Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P. (2010). A data category registry- and component-based metadata framework. Talk presented at Seventh conference on International Language Resources and Evaluation [LREC 2010]. Valletta, Malta. 2010-05-19 - 2010-05-21.

Bloom, K. (2009). Virtual Microscopy and Image Analysis. In G.L. Kumar, L. Rudbeck (Eds.), *Education guide. Immunohistochemical Staining Methods*. Carpinteria, California: Dako North America. Fifth Edition.

Cardin, V.A. (2011). 50 Years of Growth, Innovation and Leadership. Big Science Big Data Big Collaboration... …Cancer Research in Virtual Frontier. A Frost and Sullivan White Paper. Retrieved from http://www.emc.com

Chapman , J.W., Reynolds, D., and Shreeves, S.A. (2009). Repository Metadata: Approaches and Challenges. *Cataloging and Classification Quarterly,* 47 (3-4), 309 - 325.

Chung-Yueh Lien, Hsu-Chih Teng, Deng-Ji Chen, Woei-Chyn Chu, and Chia-Hung Hsiao (2009). A Web-Based Solution for Viewing Large-Sized Microscopic Images. *Journal of Digital Imaging*, Vol. 22, No. 3 (June), 275 - 285.

Cragin, M.H., Palmer, C.L., Carlson, J.R., and Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society*, 368, 4023 - 4038.

Cragin, M.H., Shreeves, S.L. (2008). Introduction: Institutional Repositories: Current State and Future. Library Trends, Volume 57, Number 2, Fall, pp. 89 - 97.

Craig, D.W., Coor, R.M., Wang, Z., Poschafl, J., Ostell, J., Feolo, M., Sherry, S.T., and Manolio, T.A. (2011). Assessing and managing risk when sharing aggregate genetic variant data. *Nature Reviews Genetics*, 12, 730-736.

Cooper, L., Carter, A., Farris, A., Wang, F., Kong, J., Gutman, D., Widener, P., Pan, T., Cholleti, S., Sharma, A., Kurc, T., Brat, D., and Saltz, J. (2012). Digital Pathology: Data-Intensive Frontier in Medical Imaging. Proceedings of the IEEE, Vol. 100, No. 4, April.

Cullen, R., Chawner, B. (2010). Institutional repositories: assessing their value to the academic community. Performance Measurement and Metrics, Vol. 11, Iss. 2, 131 - 147.

Daniel, C., Rojo, M., Bourquard, K., Henin, D., Schrader, T., Mea, V., Gilbertson, J., Beckwith, B. (2009). Standards to Support Information Systems Integration in Anatomic Pathology. *Archives Pathology and Laboratory Medicine*, Vol. 133 (February), 1841 - 1849.

Delgado, P.H., Maguitman, A.G., Ferracutti, V.M., and Herrera, L.A. (2011). Using Thematic Contexts and Previous Solutions for Maintaining and Accessing Institutional Repositories. *Journal of Computer Science and Technology*, Vol. 11, Iss. 2, 61 - 67.

Deserno, T.M., Welter, P., and Horsch, A. (2012). Towards a Repository for Standardized Medical Image and Signal Case Data Annotated with Ground Truth. *Journal of Digital Imaging*, 25, 213 - 226.

Dietrich, D. (2010). Metadata Management in a Data Staging Repository. *Journal of Library Metadata*, 10 (2-3), 79 - 98.

Doorn, P., Tjalsma, H. (2007). Introduction: archiving research data. *Archival Science*, 7, 1 - 20.

Eliceiri, K., Berthold, M., Goldberg, I., Ibáñez, L., Manjunath, B., Martone, M., Murphy, R., Peng, H., Plant, A., Roysam, B., Stuurmann, N., Swedlow, J., Tomancak, P., and Carpenter, A. (2012). Corrigendum: Biological imaging software tools. *Nature Methods*, 9, 697 - 710.

Emory University. (2013). *Pathology Analytical Imaging Standards (PAIS)*. Retrieved from http://confluence.cci.emory.edu:8090/display/PAIS/Overview

Erickson, B.J., Pan, T., Daniel, S.M. (2012). Imaging Infrastructure for Research. Part 2. Data Management Practices. *Journal of Digital Imaging*, Vol. 25, Iss. 5, 566 - 569.

Erickson, B.J., Pan, T., Daniel, S.M. (2012). White Papers on Imaging Infrastructure for Research. *Journal of Digital Imaging*, Vol. 25, Iss. 4, 449 - 453.

Erickson, B.J., Pan, T., Daniel, S.M. (2012). White Papers on Imaging Infrastructure for Research Part Three: Security and Privacy. *Journal of Digital Imaging*, Vol. 25, Iss. 6, 692 - 702.

Fear, K., and Donaldson, D.R. (2012). Provenance and Credibility in Scientific Data Repositories. *Archival Science*, 12, 319 - 339.

Flynn, S.J. (2001). The Records Continuum Model in Context and Its Implications for Archival Practice. *Journal of the Society of Archivists*, 22, 79 - 93.

Freymann, J.B., Kirby, J.S., Perry, J.H., Clunie, D.A., Jaffe, C.C. (2012). Image Data Sharing for Biomedical Research - Meeting HIPAA Requirements for De-identification. *Journal Digital Imaging*, 25, 14 - 24.

Fox, G., Hey, T., and Trefethen, A. (2011). Where does all the data come from? Bloomington, IN, Indiana University. Retrieved from http://grids.ucs.indiana.edu

Garcia, M.S., and Trefethen, A.E. (2012). A Virtual Research Environment for Cancer Imaging. VRE-CI Final Report. Retrieved from http://www.jisc.ac.uk

Goldberg, I., Allan, C., Burel, J., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P., Swedlow, J. (2005). The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biology*, 6 (5).

Grace, S., Knight, G., Montague, L. (2009). Investigating the Significant Properties of Electronic Content over Time – InSPECT Final Report. Retrieved from http://www.significantproperties.org.uk

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. *Cataloging and Classification Quarterly*, Vol. 40, No. 3 - 4, 17 - 36.

Greenberg, J. (2003). Metadata generation: process, people, tools. *The Bulletin of the American Society for Information Science and Technology*, December-January.

Greenberg, J., White, H.C., Carrier, S., and Scherle R. (2009). A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata*, 9 (3-4), 194 - 212.

Griffiths, A. (2009). The Publication of Research Data: Researcher Attitudes and Behaviour. *The International Journal of Digital Curation*, Vol. 4, No. 1, 46 - 56.

Groenewald, R., Breytenbach, A. (2011). The Use of Metadata and Preservation Methods for Continuous Access to Digital Data. *Electronic Library*, vol. 29 (2), 236 - 248.

Gubrium, J.F., Holstein, J.A., Marvasti, A.B., and McKinney, K.D. (Ed.). (2012). *The SAGE Handbook of Interview Research. The Complexity of the Craft.* Thousand Oaks, California: SAGE Publications, Inc.

Hamilton, R.J., and Bowers, B.J. (2006). Internet Recruitment and E-Mail Interviews in Qualitative Studies. *Qualitative Health Research*, Vol. 16 (6), 821 - 835.

Hart, K. (2006). Inbox Journalism. *American Journalism Review*, December-January. Retrieved from http://www.ajr.org

Heery, R. and Anderson, S. (2005) Digital repositories review. UKOLN and AHDS Report. Retrieved from http://www.jisc.ac.uk

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, Vol. 57, No. 2, 280 - 299.

Hey T, Hey J. (2006). E-Science and Its Implications for the Library Community. *Library Hi Tech*, 24, 515 - 528.

Hockx-Yu, H. (2006). Digital Preservation in the Context of Institutional Repositories. *Program: Electronic Library and Information Systems*, Vol. 40 Iss. 3, 232 - 243.

Hodge, G.M. (2000). Best Practices for Digital Archiving. An Information Life Cycle Approach. *Digital Library Magazine*, Vol. 6, No. 1. Retrieved from http://www.dlib.org/dlib/january00/01hodge.html

Hunt, N., McHale, S. (2007). A Practical Guide to the E-Mail Interview. *Qualitative Health Research*, Vol. 17 (10), 1415-1421.

Ixchel, F. M., and Yakel, E. (2011). Significant Properties as Contextual Metadata. *Journal of Library Metadata*, 11 (3-4), 155-165. Preprint. Retrieved from http://www.oclc.org

Jean, B., Rieh, S.Y., Yakel, E., Markey, K. (2011). Unheard Voices: Institutional Repository End-Users. *College and Research Libraries*, Vol. 72, No. 1, 21 - 42.

Karasti, H., Baker, K.S., and Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work*, Vol. 15, Iss. 4, 321 - 358.

Kim, H.H., Kim, Y.H. (2008). Usability Study of Digital Institutional Repositories. *The Electronic Library*, Vol. 26, Iss. 6, 863 - 881.

Korenblum, D., Rubin, D., Napel, S., Rodriguez, C., and Beaulieu, C. (2011). Managing Biomedical Image Metadata for Search and Retrieval of Similar Images. *Journal of Digital Imaging*, 24(4), 739 - 748.

Kowalczyk, S.T. (2011). *E-Science Data Environments: A View from the Lab Floor.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. UMI No. 3491487.

Langer, S. G. (2011). Challenges for Data Storage in Medical Imaging Research. *Journal of Digital Imaging*, Vol. 24, No. 2, 203 - 207.

Laurinaviciusa, A., Laurinavicienea, A., Daseviciusa, D., Elied, N., Plancoulained, B., Bord, C., Herlin, P. (2012). Digital Image Analysis in Pathology: Benefits and

Obligation. *Analytical Cellular Pathology*, 35, 75–78. DOI 10.3233/ACP-2011-0033

Laxminarsaiah, A., Rajgoli, I. U. (2007). Building Institutional Repository: an Overview. *OCLC Systems and Services*, Vol. 23, Iss. 3, 278 - 286.

Lee, C. A. (2011). A framework for contextual information in digital collections. *Journal of Documentation*, Vol. 67, Iss. 1, 95 - 143.

Levy, M.A., Freymann, J.B., Kirby, J.S., Fedorov A., Fennessy, F.M., Eschrich, S.A., Berglund, A.E., Fenstermacher, D.A., Tan, Y., Guo, X., Casavant, T.L., Brown, B.J., Braun, T.A., Dekker, A., Roelofs, E., Mountz, J.M., Boada, F., Laymon, C., Oborski, M., Rubin, D.L. (2012). Informatics Methods to Enable Sharing of Quantitative Imaging Research Data. *Magnetic Resonance Imaging*, 30, 1249 - 1256.

Linkert, M., Rueden, C., Allan, C., Burel, J., Moore, W., Patterson, A., Loranger, B., Moore, J., Neves, C., MacDonald, D., Tarkowska, A., Sticco, C., Hill, E., Rossner, M., Eliceiri, K., and Swedlow, J. (2010). Metadata Matters: Access to Image Data in the Real World. *Journal of Cell Biology*, Vol. 189, 5777 - 5782.

Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL*, 226, 1 - 7.

Martinez-Uribe, L. (2007). Digital Repository Services for Managing Research Data: What Do Oxford Researchers Need? *IASSIST Quarterly*, Fall/Winter.

Martinez-Uribe, L., and Macdonald, S. (2009). User Engagement in Research Data Curation. *Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science,* Vol. 5714, 309 - 314.

Mayernik, M., Choudhury, S., DiLauro, T., Duerr, R., Metsger, E., Pralle, B., and Rippin, M. (2012). The Data Conservancy Blueprint for Data Management. Retrieved from https://dataconservancy.org

McGuire, A.L., Hamilton, J.A., Lunstroth, R., McCullough, L.B., and Goldman, A. (2008). DNA Data Sharing: Research Participants' Perspectives. *Genetics in Medicine*, 10 (1), 46 - 53.

McKemmish, S. (1997). Yesterday, Today and Tomorrow: A Continuum of Responsibility. Proceedings of the Records Management Association of Australia, 14th National Convention, 15–17 September 1997, RMAA Perth.

Meho, L.I. (2006). E-mail Interviewing in Qualitative Research: A Methodological Discussion. *Journal of the American Society for Information Science and Technology*, 57 (10), 1264 - 1295.

Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., and Jamali, H. R. (2012). Digital Repositories Ten Years On: What Do Scientific Researchers Think of Them and How Do They Use Them? *Learned Publishing*, Vol. 25, No. 3, 195 - 206.

Normore, L. F., Tebo, M. E. (2011). Assessing User Requirements for a Small Scientific Data Repository. ASIST, October 9-13, New Orleans, LA, USA.

Onsrud, H., and Campbell, J. (2007). Big Opportunities in Access to “Small Science” Data. *Data Science Journal*, Vol. 6, Open Data Issue.

Osswald, A. (2008). E-science and Information Services: A Missing Link in the Context of Digital Libraries. *Online Information Review*, Vol. 32, Iss. 4, 516 - 523.

Park, S., Pantanowitz, L., Parwani, A.V. (2012). Digital Imaging in Pathology. *Clinics in Laboratory Medicine*, 32(4), 557 - 584.

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., and Carlson, D. (2011). A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science. *Journal of Information Science*, 37 (555).

Plale, B.; Kouper, I.; Seiffert, K.; Konkiel, S.R. (2013). Repository of NSF-funded Publications and Related Datasets: "Back of Envelope" Cost Estimate for 15 years. Technical Report. Working Paper. Indiana University. Retrieved from http://hdl.handle.net/2022/16599

Rambo, N. (2009). E-science and biomedical libraries. *Journal Medical Library Association*, 97(3).

Ray, J. (2012). The Rise of Digital Curation and Cyberinfrastructure: From Experimentation to Implementation and Maybe Integration. *Library Hi Tech*, Vol. 30, Iss. 4, 604 - 622.

Renear, A.H., Sacchi, S., and Wickett, K.M. (2010). Definition of Dataset in the Scientific and Technical Literature. ASIST 2010, October 22–27, Pittsburgh, PA, USA. Retrieved from http://mail.asist.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/240_Final_Submission.pdf

Riding the Wave. How Europe Can Gain From the Rising Tide of Scientific Data. Final Report of the High Level Expert Group on Scientific Data. A Submission to the European Commission. October, 2010.

Schröder, P. (2007). Possible Downsides to Data Sharing in the Research Commons: Assets and Liabilities, Opportunities and Risks. *Data Science Journal*, Vol. 6, Open Data Issue.

Seto, B., and Luo, J. (2007). Biomedical Data Sharing, Security and Standards. *Data Science Journal*, Vol. 6, Open Data Issue.

Spidlen, J., Shooshtari, P., Kollmann, T.R., and Brinkman, R.R. (2011). Flow cytometry data standards. *BMC Res Notes*, Mar 7; 4:50.

Stuurman, N., and Swedlow, J. (2012). Software Tools, Data Structures, and Interfaces for Microscope Imaging. Cold Spring Harbor Protocols.

Swedlow, J., and Eliceiri, K. (2009). Open Source Bioimage Informatics for Cell Biology. Trends Cell Biology, 19 (11-3), 656 - 660.

Swedlow, J., Goldberg, I., Eliceiri, K., and the OME Consortium. (2009). Bioimage Informatics for Experimental Biology. *The Annual Review of Biophysics*, Vol. 38, 327 - 346.

To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. (2008). Report commissioned by the Research Information Network (RIN).

White, H.C. (2010). Considering Personal Organization: Metadata Practices of Scientists. *Journal of Library Metadata*, 10, 156 - 172.

Willis, C., Greenberg, J., and White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Data. J*ournal of the American Society For Information Science and Technology*, 63 (8), 1505 - 1520.

Wilson, A. (2010). How Much Is Enough: Metadata for Preserving Digital Data. *Journal of Library Metadata,10 (2-3),* 205 - 221.

Wilson, M. C., Jantz, R. C. (2011). Building Value-added Services for Institutional Repositories (IRs): Modeling the Rutgers Experience. International Federation of Library Associations Social Science Libraries Section, Satellite Conference Social

Science Libraries: A Bridge to Knowledge for Sustainable Development. Biblioteca Nacional de Cuba José Martí, Havana, Cuba 8 - 10 August.

Yagi, Y., Gilbertson, J. (2005). Digital Imaging in Pathology: the Case for Standardization. *Journal of Telemedicine and Telecare*.

Yeates, R. (2003). Institutional repositories. *VINE*, Vol. 33, Iss. 2, 96 - 101.

Appendix I. Survey

#	Questions	Answers
1	Please describe your research area/focus and the methodology you use.	
2	How many researchers work in your lab?	
3	Please name the file formats that you regularly use in your research (e.g., .tif, .jpg, .doc, .xls, etc).	
4	Do you use commercial or lab derived software to generate your data? Please explain.	
5	Do you use commercial or lab derived software to analyze your data? Please explain.	
6	What is the average size of the data sets you generate during an experiment/study?	
7	Do you have a set of formalized quality control criteria to ensure that the integrity of the data and files that you have remain intact and accessible? If so, please describe.	
8	Do you describe your experiments, the conditions used during the experiments, the parameters used during data acquisition, and the changes that occur when you process the data? If so, is this description in electronic or paper form?	
9	Do you have special lab templates for the description of your research protocols and analysis?	

10	How many descriptive elements do you use to describe an individual data set? A descriptive element could be time, date, antibodies used, hypothesis being tested, etc.	
11	How many descriptive elements do you use to describe your data collections?	
12	Which metadata standards do you use?	
13	Do you generate metadata automatically? If so, what is the process?	
14	In what kind of format/file format do you preserve metadata?	
15	Who do you think should be responsible for metadata creation (e.g., researcher, data professional, librarian or someone else)?	
16	Do you have experience with depositing research data into a digital repository, data warehouse?	
17	Where should files be preserved/archived (e.g., on lab storage devices, in institutional repositories, or in community supported research collections)?	
18	What kind of data should be preserved in a research digital repository (published, unpublished, raw data, aggregated data files)?	
19	Do you think it is important to publish explanations as to how researchers generated data (instruments, experiments) and what methodologies researchers used? Please explain.	

20	Who should pay for the cost associated with storing the data/files in a repository? That is, should it be the researchers who publish the data, the researchers who access the digital repositories, commercial companies (e.g., pharmacies) who want to re-analyze research data, or should all data be completely free for everybody and supported by the local, state, or federal government?	
21	What kinds of services do you think digital repositories should have? Would you use those services for a fee?	
22	What kind of digital repositories should be inter-institutional, cross institutional, national?	

Appendix II. Consent Form

University of North Carolina at Chapel Hill
Consent to Participate in a Research Study
Adult Participants

Consent Form Version Date: __April 27, 2013__
IRB Study # 13-1159
Title of Study: Scientific Digital Data Repositories: Needs and Challenges for Cancer Researchers
Principal Investigator: Maria Ryshkevich
Principal Investigator Department: School of Information and Library Science, University of North Carolina at Chapel Hill
Principal Investigator Phone number: (919) 923-9155
Principal Investigator Email Address: mchastai@email.unc.edu
Faculty Advisor: Dr. Helen Tibbo
Faculty Advisor Contact Information: tibbo@ils.unc.edu
School of Information and Library Science, University of North Carolina at Chapel Hill
216 Lenoir Drive • CB #3360 • 100 Manning Hall, Chapel Hill, NC 27599-3360

What are some general things you should know about research studies?
You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, without penalty.

Research studies are designed to obtain new knowledge. This new information may help people in the future. You may not receive any direct benefit from being in the research study. There also may be risks to being in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study.

You will be given a copy of this consent form. You should ask the researchers named above, or staff members who may assist them, any questions you have about this study at any time.

What is the purpose of this study?
The purpose of this study is to understand the information management needs of cancer researchers who use light and confocal microscopy to obtain knowledge about cancer on a molecular level. The study will focus on how these researchers are preserving their data and using digital repositories so as to provide future access to this data. Data will be collected via an email survey instrument.
You are being asked to be in the study because you generate research data on a molecular level using light and confocal microscopy and you had been or currently are a Principal Investigator (PI) and have 2 or more people in your research lab.

How many people will take part in this study?
A total of approximately 15-20 people at number institutions will take part in this study, including approximately 3 people from this institution.

How long will your part in this study last?
The survey will take you approximately one hour to fill out.

What will happen if you take part in the study?
After agreeing to be part of this study, you will be sent a questionnaire to fill out and return by email. The questionnaire will contain questions concerning the type of research you do, what kind of file formats you use, how you describe your experiments, whether you use metadata, how you store your data, and your thoughts concerning depositing your research data in data repositories. You may skip any questions that make you uncomfortable.

What are the possible benefits from being in this study?
Research is designed to benefit society by gaining new knowledge. You will not benefit personally from being in this research study.

What are the possible risks or discomforts involved from being in this study?
There will be only minimal risks associated with this project. Survey responses will be stripped from any identifying information upon receipt of the data. Only de-identified data will be stored. There may be uncommon or previously unknown risks. You should report any problems to the researcher.

What if we learn about new findings or information during the study?
You will be given any new information gained during the course of the study that might affect your willingness to continue your participation.

How will information about you be protected?
To ensure that no link exists between you and your completed survey once your message is received via email, your responses will be separated from the message, and your email will be promptly deleted and purged from the email system so as to preserve your confidentiality. The responses to each question will be checked for any identifying information which will be removed, again to assure confidentiality. The responses will be compiled and analyzed along with the data from the other participants in the study. During the research process only principal investigator will have access to individually identifiable data.

Participants will not be identified in any report or publication about this study. Although every effort will be made to keep research records private, there may be times when federal or state law requires the disclosure of such records, including personal information. This is very unlikely, but if disclosure is ever required, UNC-Chapel Hill will take steps allowable by law to protect the privacy of personal information. In some cases, your information in this research study could be reviewed by representatives of the

University, research sponsors, or government agencies (for example, the FDA) for purposes such as quality control or safety.

<u>What if you want to stop before your part in the study is complete?</u>
You can withdraw from this study at any time, without penalty.

<u>Will you receive anything for being in this study?</u>
You will not receive anything for taking part in this study.

<u>What if you have questions about this study?</u>
You have the right to ask, and have answered, any questions you may have about this research. If you have questions about the study (including payments), complaints, concerns, or if a research-related injury occurs, you should contact the researchers listed on the first page of this form.

<u>What if you have questions about your rights as a research participant?</u>
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject, or if you would like to obtain information or offer input, you may contact the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

<u>Participant's Agreement</u>:

I have read the information provided above. I have asked all the questions I have at this time. I voluntarily agree to participate in this research study.

____ ____________
Signature of Research Participant Date

____
Printed Name of Research Participant

____ ____________
Signature of Research Team Member Obtaining Consent Date

____
Printed Name of Research Team Member Obtaining Consent