2011

# Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder

Kathryn Lage
*University of Colorado, Boulder*

Barbara Losoff
*University of Colorado Boulder*, barbara.losoff@Colorado.EDU

Jack M. Maness
*University of Denver*, jack.maness@du.edu

## Recommended Citation

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the university of colorado boulder. portal: Libraries and the Academy, 11(4), 915-937. https://doi.org/10.1353/pla.2011.0049

# Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder

**Kathryn Lage, Barbara Losoff, and Jack Maness**

**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."[1] 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.

## Introduction

This paper investigates the current state of data curation activities and existing curation support for faculty and graduate students in the sciences at the University of Colorado Boulder. Data curation is "the management and preservation of digital data over the long-term"[2] and is a growing topic in library science. Grounded in the growing body of literature of related surveys and needs assessments, this case study explores needs, although it is not a detailed needs assessment. Rather, its intent is to gauge researcher receptivity to library involvement in scientific data curation so that partnerships between the library and scientists can be strategically developed. Conclusions drawn from this research are applicable to any academic library seeking to provide data curation support to scientific researchers.

The authors interviewed researchers and conducted a qualitative analysis of the interviews in order to develop personas. Personas allow for anonymization and aggre-

*portal: Libraries and the Academy*, Vol. 11, No. 4 (2011), pp. 915–937.

gation of ethnographic research, in this case, interviews. They are used to personalize and strengthen the tie between a systems designer and the target user. The study culminated in the development of eight personas that embody the aggregated attributes of faculty and graduate student researchers who were interviewed. The authors analyzed the interviews and created the personas to represent the range of attitudes and needs regarding the type of datasets created, existing data storage and maintenance support, disciplinary culture or personal feelings on data sharing, and receptivity to the library's role in data curation.

Through the creation of personas the authors intend to keep researchers' needs front and center in the design of a library-led institutional repository for research data. The authors hope to identify those scientists who are most receptive to the library's involvement in research data curation so they may be approached and partnered with strategically. Many researchers already have disciplinary repositories, departmental networks, or systems within their laboratories that serve as mechanisms by which data can be preserved. Others are protective of their data, wary of outside access. Purposefully partnering with researchers who are receptive to the idea of library involvement in such data curation initiatives, then, is wise. The authors believe that the personas developed in this study will be recognizable to librarians at other institutions, rendering the findings generalizable and open to quantitative validation.

## Literature Review

This review focuses on the primary literature that contextualizes this study: data curation, which may be considered a sub-discipline of e-science; and the development and use of personas. In both areas the review provides a brief background of foundational writings and focuses on their application to libraries and librarians.

### Data Curation

Increasingly, libraries are expected to play a role in data curation. This is a result of what Tony and Jessie Hey term the "e-science revolution," producing data at increasing orders of magnitude, and the U.S. government funding policies regarding data management and access.[3] In 2007 the National Science Foundation (NSF) mandated that "all science and engineering data generated with NSF funding must be made broadly accessible and usable while being suitably protected and preserved."[4] Libraries play an integral part of the NSF data curation model and are viewed as trusted repositories, according to Lucy Nowell, "preserving documents that have been the foundation of civilizations."[5] In fact, the NSF, in their solicitation and funding guidelines for Sustainable Digital Preservation and Access Network Partners (DataNet) grants, describes a new organization that "will integrate library and archival sciences, cyberinfrastructure, computer and information sciences and domain science expertise."[6] Despite the model proposed by the NSF, the question of whether libraries can play a role in data curation and archiving has continued to be a topic of debate.[7] More recently, however, the literature has moved away from articles that question the role of librarians in data curation, toward articles reporting on active data curation or existing data repositories. Imperatives such as Joyce L. Ogburn's statement “librarians will have to embrace the role of data curator to remain relevant and vital to our scholars”[8] and Tracy Gabridge’s position that“[data liaison services] are a major component of libraries’ future”[9] have become the more common refrain.

In an effort to define cyberinfrastructure for librarians and explain how it is applicable to libraries, Anna Gold from the Massachusetts Institute of Technology (MIT) offered a two-part primer in *D-Lib Magazine* in 2007. In part one, she provides an overview of the issues surrounding cyberinfrastructure, and in part two, lists actions for developing a role in data librarianship. Among the possible roles for librarians, she includes social science data services, geo-referenced data services (GIS), and bioinformatics. Although Gold describes these roles for librarians as “well-defined,” she also questions “the extent of the roles libraries can play in relationship to data archiving.”[10]

Nonetheless, many academic libraries are developing or have launched data repositories, though the extent of their involvement in data archiving remains to be seen. Academic libraries exploring data curation include Purdue University, Georgia Institute of Technology (Georgia Tech), Cornell University, and Johns Hopkins University. The Purdue Libraries investigated discipline-based data workflow through faculty workshops, seminars, and surveys. Engaging the faculty on the topics of data discovery, management, and organization, the librarians found that researchers uniformly expressed a need for organizing, describing, managing, archiving, and accessing data.[11] The Purdue archiving project resulted in *Purdue e-Data* which is still under development. The librarians, in collaboration with faculty from engineering, agronomy, physics, and earth & atmospheric sciences, are, according to Michael Witt, working “to populate new data collections and [to] experiment with them.”[12]

**Engaging the faculty on the topics of data discovery, management, and organization, the librarians found that researchers uniformly expressed a need for organizing, describing, managing, archiving, and accessing data.**

The Georgia Tech Library followed a similar path to Purdue’s, based on what Tyler Walters describes as the “clear need for data curation put forth by researchers.” Georgia Tech convened a *Data Curation Workgroup* to interview select groups of neuroscientists and bioscientists regarding data retention, data sharing, and storage. Walters’ paper outlined Georgia Tech’s experience and offered a four-point model to guide librarians in developing data curation programs. Walters stressed that grant funds are necessary for data curation and that locating the ”early-adopting researcher with whom to explore data curation approaches is critical.”[13]

Gail Steinhart and John Saylor describe a pilot program at Cornell University’s Albert R. Mann Library that “test[s] the feasibility of a local ‘staging’ repository to support data sharing among research collaborators while research is in progress, and to provide tools and support to publish data to permanent disciplinary or institutional repositories.”[14] DataStaR (http://datastar.mannlib.cornell.edu/) is a data staging repository funded in part by the NSF in order to provide a means for recruiting digital data into a domain-specific repository or an institutional repository. This collaborative effort between researchers and librarians has already produced more than twenty published data sets (including metadata); however, it remains to be seen whether a non-permanent repository, such as DataStaR, will be the most successful model for data curation.

In October 2009 the NSF awarded the Sheridan Libraries at Johns Hopkins University a five-year, $20 million dollar grant to build a research data infrastructure for managing digital information created by the life, earth, and social sciences.[15] This Data Conservancy grant is a collaborative project with Sayeed Choudhury, associate dean of the libraries, serving as principal investigator, along with faculty members from Johns Hopkins University and individuals from other institutions.

Indeed, collaboration, a key component of the Data Conservancy project, is reflected elsewhere in the literature. The practice of librarianship, described by Gold, is "rooted in the management and 'delivery' of relationships" and "data is an encoding of relationships in the world, whether those relationships involve instruments, physical phenomena, social entities, measurements, time, place, or other intellectual constructs."[16] Jennifer Haas and Sharon Murphy remind librarians that "we must know the data options and obligations of the disciplines we serve and be ready to facilitate this communication."[17] According to an ARL Task Force, librarians need to "develop a deep understanding of content users and researchers needs,"[18] and to cultivate, as Karla Hahn says, "new forms of relationship building."[19]

To further develop these relationships, Witt and others created a *Data Curation Profile*, "to provide detailed information on particular data forms that might be curated by an academic library" and to "address a perceived shortage of robust models for the systematic description of datasets for sharing and curation."[20] The *Data Curation Profile* also functions as a tool to evaluate data curation from the perspective of the producers. Such a profile may be considered a tactical model by which a library could develop a process to archive and describe data sets. The literature suggests, then, that libraries are a critical partner in a national effort toward the curation of scientific data, not the sole responsible party.

**The *Data Curation Profile* also functions as a tool to evaluate data curation from the perspective of the producers. Such a profile may be considered a tactical model by which a library could develop a process to archive and describe data sets.**

### Personas

The research described in this paper attempts to broaden the discussion regarding scientific data curation and libraries through the development of personas. First suggested by Alan Cooper as a mechanism for product development, personas are fictionalized aggregates of actual potential users of a product. Personas provide anonymous findings that can be more broadly generalized. Even though they are fictitious, personas are intended to create personal empathy between a designer and the target user so the needs and goals of the user are holistically and continuously considered during the design process. The intent is that all design decisions, from content to functionality, are driven not by the convenience of the designer, but by those of the persona. Building on Cooper's original thesis, John Pruitt and Tamara Adlin have explored the psychological evidence supporting personas and have advanced the design applications.[21] Though empirical findings validating the approach are not voluminous, the work of Frank Long suggests that using personas results in "designs with better usability attributes" and that there is "some objective evidence that using personas does work."[22]

Personas are created in many ways, but should always be driven by the data collected from actual or potential users. One of the most common ways of collecting data is to interview many users and find shared design needs among them. Users with several mutual needs are then conflated into a single persona, and that persona is given a name, biographical information, a description of his or her intent in using the product, and often a facial image that can enhance empathy. There can be as many or as few personas as the interviews suggest, and specific demographics portrayed in the persona are not meant to represent the demographics of the interviewees. The designer is still an entity in this part of the analysis, possibly introducing any misinterpretations he or she may have of user needs into the personas. Some researchers are exploring ways to automate this process, thereby further removing the designer from the creation of the personas.[23] This research is just emerging, however, and the most common method still involves designers analyzing and interpreting user data. The study described in this article utilizes this more conventional method. Using personas in the design of library websites is not uncommon.[24] And in a previous research project, librarians and researchers at the University of Colorado Boulder used this process to identify the general needs and goals of institutional repository users.[25] Findings from that project help inform the design of repositories and associated policies regarding what materials should be collected. Expressed needs by interviewees related to data, however, were infrequent at best, and the research questions did not address ancillary challenges related to intellectual property, disciplinary culture, and storage and format problems, and did not focus on scientists. The use of personas to determine needs with respect to scientific data curation, then, particularly library involvement in that curation, is novel. The creation of personas in this context is not a continuation of previous studies, but is complementary to them.

**Users with several mutual needs are then conflated into a single persona, and that persona is given a name, biographical information, a description of his or her intent in using the product, and often a facial image that can enhance empathy.**

## Methods

The authors recruited participants representing the broad range of scientific disciplines found at the University of Colorado Boulder, identified through departmental rosters. Additionally, some participants were recruited based on previous connections with either the authors or the library, or through referrals. Authors contacted participants via email explaining that the authors were investigating research data creation and curation at CU Boulder and inviting the researchers to participate in a 15-30 minute interview.

An official letter of invitation was attached to the email, which described the research in more detail. The authors framed the research in terms of current issues regarding data curation locally and nationally: the University Libraries' planned participation in a cooperative institutional repository, the National Science Foundation's (NSF) mandate that data generated with NSF funding "be made broadly accessible and usable, while being suitably protected and preserved"[26] and the NSF's suggestion that libraries and archives play a part in carrying out that mandate.[27] (See Appendix 1 for the full text of the letter of invitation.)

The authors contacted thirty-five researchers, following up the email invitation with phone calls or additional emails as needed. The authors interviewed a total of twenty-six researchers—a response rate of 74 percent. Nineteen of the interviewees were faculty members, three were doctoral students, and four were research associates. The participants represented nine departments within the sciences at the University of Colorado Boulder: the Departments of Aerospace Engineering Sciences, Civil, Environmental and Architectural Engineering, Mechanical Engineering, Computer Science, Ecology and Evolutionary Biology, Chemistry and Biochemistry, Molecular, Cellular, and Developmental Biology, Geography and Geological Sciences. Additionally, researchers interviewed represented three research centers affiliated with the campus. The researchers' subject areas included evolutionary biology, organic chemistry, chemical genetics, climatology, geographic information science, environmental engineering, construction management, nanotechnology, and aeronautical engineering.

The authors asked a fixed set of nine questions, with follow-up as necessary. Participants were asked to briefly describe their research, including the type of data created; how the data was stored; how the data was accessed and by whom; if there were procedures for data preservation; and whether storage space was an issue. The final two questions gauged the researchers' receptivity to the library's involvement in data curation service: first asking the question in broad terms and then asking the researchers to rank their level of interest on a scale of 1-5. Most interviews went well beyond the 15-30 minutes stated in the letter of invitation, with many lasting 45 minutes to one hour. (See Appendix 2 for the full text of the survey questions.)

**Participants were asked to briefly describe their research, including the type of data created; how the data was stored; how the data was accessed and by whom; if there were procedures for data preservation; and whether storage space was an issue.**

In order to facilitate discussion, all of the interviews were conducted in person at a location convenient for the researcher. Not all the authors were able to attend each interview; however, the authors who were present took detailed notes. After the interviews were completed, the authors transferred the information into a grid on a large chalkboard as a means to easily identify the themes or patterns. The authors used a color-coded schematic to highlight the common themes from the researchers' answers. This technique of grouping answers together in the grid, while highlighting research domain, exposed similarities in types of data created, existing storage and maintenance, accessibility of the data, procedures for data preservation, and receptivity to a library role in data curation.

The common aspects that emerged from the grid were used to develop the personas. Initially, the authors developed twelve personas, which were later reduced to a more distinct eight. These eight personas represent the aggregated attributes of the researchers.

As with any methodology, the development of personas has its limitations. Like any ethnographic research, data gathered through interviews is filtered through the lens of the authors themselves. While many of the questions used in the development of these personas were objective (i.e. "how is the data stored and accessed?"), others were more subjective (i.e. "is storage space problematic?" and, most significantly, the last two questions regarding receptivity to a library-run data repository).

It is important to note that, as Steve Mudler and Ziv Yaar point out, the creation of personas using interviews "consists of interacting with a small number of users (10-20) to get new ideas or uncover previously unknown issues. [It] doesn't prove anything … but it's very valuable at uncovering insights that you can then test and prove."[28] In other words, personas used in this context will not tell a designer or liaison that "37 percent of faculty would contribute data to an IR," but instead will bring potential contributors' needs alive. Personas created in this project can be used in exactly such a manner; as a starting point for more quantitative analyses of surveys and participation rates among scientists in library-led data curation projects. Results can help inform library administrators so they may effectively position data services on campus, and also inform science librarians about the issues and challenges related to library-provided data curation.

In addition, the final interview question gauging receptivity asks the researcher to make a decision regarding a repository that does not exist; there is no interface and no supporting database, no deposit or access policies. The researcher cannot be expected to fully envision the repository, and this gap could have affected receptivity. Gauging receptivity before product development, however, is an important step in user-centered design, and the authors developed the personas presented here using the commonly accepted practices in the field of human-computer interaction.

## Results

The types of data created and collected range from paper notebooks, to digital ASCII files created from field instruments or software modeling programs, to large video and image files. Much of the data are software or instrument-specific. Digital data are stored on researchers' hard drives, on laboratory servers, in removable storage devices such as CDs, and in national and international discipline-specific repositories. One interviewee uses *Google Documents* as a cloud storage solution, while other researchers keep their data in paper notebooks or files, usually stored in the researcher's office, but on some occasions in the researcher's home.

Most researchers who participated in this study identified their research data as non-public (20 out of 26). What this means as a practical matter varied. Some researchers share their data within their lab or with other collaborators, while others do not share their data at all. Many identified themselves as "gatekeepers," responding that their data is not public, but they would share their data with another researcher if they considered it appropriate. Only six researchers (roughly 25 percent) identified their data as being public. Issues of confidentiality, classified research, or privacy clauses that prevented sharing of data were frequent considerations by researchers, crossing disciplinary boundaries.

**Issues of confidentiality, classified research, or privacy clauses that prevented sharing of data were frequent considerations by researchers, crossing disciplinary boundaries.**

Most interviewees stated that while digital storage space is not an issue, server maintenance and management are on-going problems. In some situations, these problems are significant and presented considerable costs to the researchers in both capital and labor. These interviewees view data maintenance as an additional responsibility that takes them away from their research. Problems associated with data management include the transient nature of graduate student network managers, the time-intensive aspect of server maintenance, and the lack of departmental expertise in managing the data.

Many researchers had curation plans in place for much of their data, but also had subsets of orphan data. These data were either of a different type, fell outside the main scope of their research, or for other reasons were not treated in the same manner, and therefore, the researchers had no curation or maintenance plan. For many of these researchers, these data were not foremost in their mind, but when asked about them, they realized they did need assistance; they were forgotten datasets from their research.

Few interviewees had departmental procedures for data preservation. However, some researchers participated in disciplinary-based repositories that supported long-term storage of their data while others had regular data back-up procedures for their lab.

Some researchers were interested in a library-sponsored repository simply because they had no other assistance in managing their data. Other researchers had restrictions on data sharing and even data management by an outside party due to research subject privacy, confidentiality clauses for contracted research, or national security concerns. As a result, those researchers were hesitant or completely unable to turn over curation of their data to an outside party.

The authors found that a researcher's receptivity to a library role in data curation did not necessarily correspond with the level of need for assistance or existence of obstacles. Rather, the individual researcher's disciplinary culture or philosophy regarding data sharing and collaborative projects affected his receptivity. Some researchers believed that sharing data or making data accessible for possible use in the future was, simply, a good thing to do. One interviewee responded that, "there is no sense in collecting data if it can't be used [by other researchers]." Others did not entertain the possibility of sharing data, in spite of a need for data management support or the absence of restrictions on their data. In addition, many interviewees expressed interest in a library role in data curation if the data were easy to access and if the researchers were

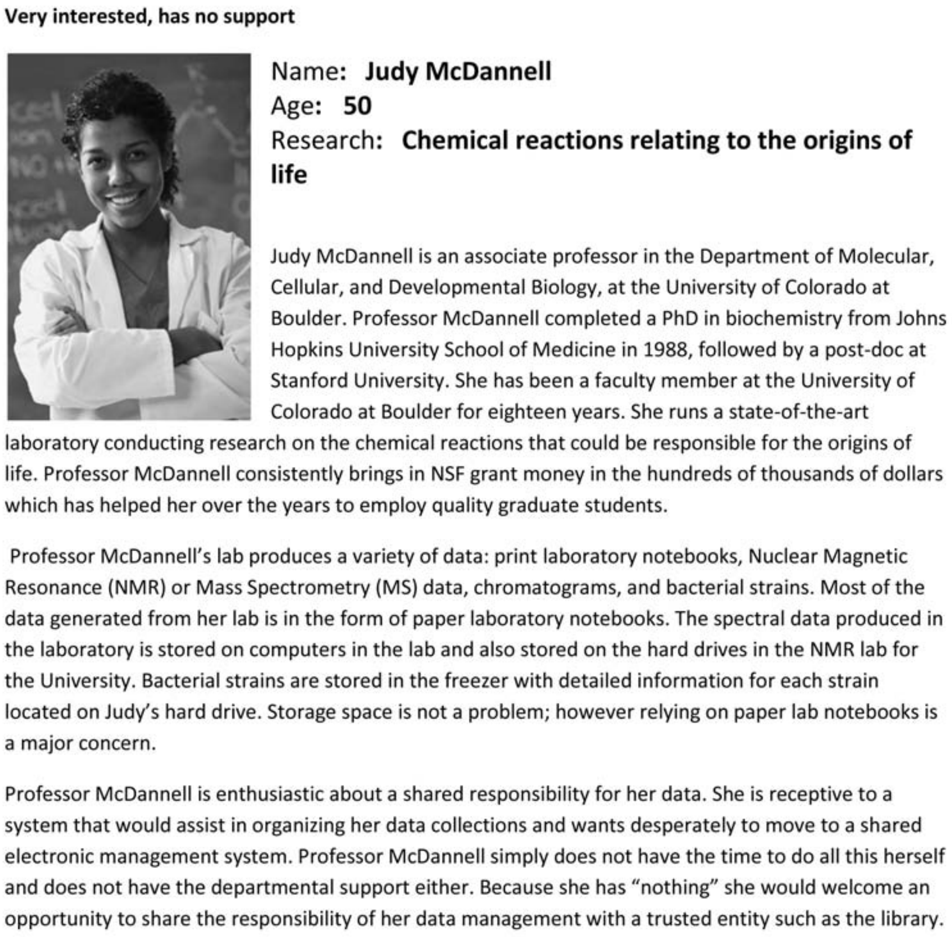
**Ease of use was by far the most-stated requirement by the researchers interested in a library-run repository.**

able to maintain a level of control over who had access to the data. Ease of use was by far the most-stated requirement by the researchers interested in a library-run repository.

Out of these themes, the authors developed eight personas. Of the twenty-six interviewees, three fell outside the scope of a library-run repository because they are data or collection managers for departments or large repositories. An additional three interviewees are also actively involved in data curation. The authors did not create personas to represent these interviewees.

The eight personas are:

Judy McDannell, "Very interested, has no support"



**Very interested, has no support**

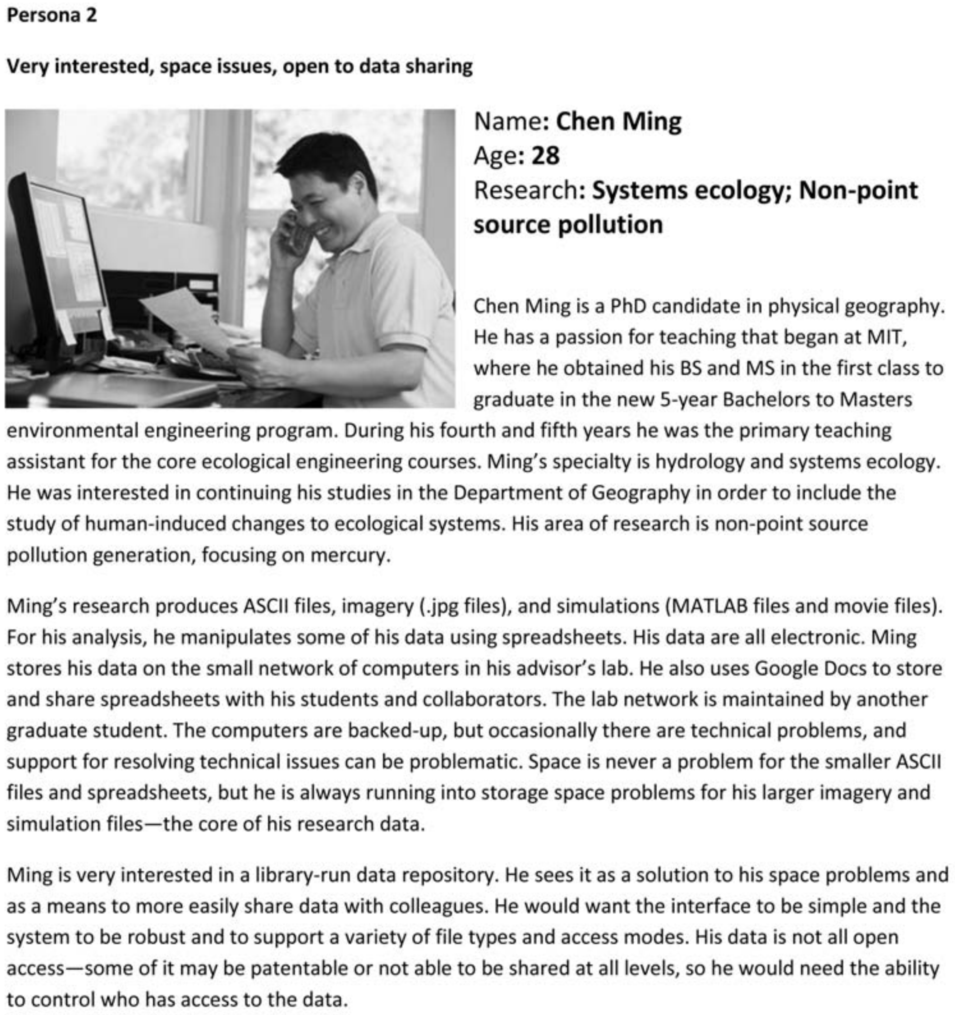
Name: **Judy McDannell**
Age: **50**
Research: **Chemical reactions relating to the origins of life**

Judy McDannell is an associate professor in the Department of Molecular, Cellular, and Developmental Biology, at the University of Colorado at Boulder. Professor McDannell completed a PhD in biochemistry from Johns Hopkins University School of Medicine in 1988, followed by a post-doc at Stanford University. She has been a faculty member at the University of Colorado at Boulder for eighteen years. She runs a state-of-the-art laboratory conducting research on the chemical reactions that could be responsible for the origins of life. Professor McDannell consistently brings in NSF grant money in the hundreds of thousands of dollars which has helped her over the years to employ quality graduate students.

Professor McDannell's lab produces a variety of data: print laboratory notebooks, Nuclear Magnetic Resonance (NMR) or Mass Spectrometry (MS) data, chromatograms, and bacterial strains. Most of the data generated from her lab is in the form of paper laboratory notebooks. The spectral data produced in the laboratory is stored on computers in the lab and also stored on the hard drives in the NMR lab for the University. Bacterial strains are stored in the freezer with detailed information for each strain located on Judy's hard drive. Storage space is not a problem; however relying on paper lab notebooks is a major concern.

Professor McDannell is enthusiastic about a shared responsibility for her data. She is receptive to a system that would assist in organizing her data collections and wants desperately to move to a shared electronic management system. Professor McDannell simply does not have the time to do all this herself and does not have the departmental support either. Because she has "nothing" she would welcome an opportunity to share the responsibility of her data management with a trusted entity such as the library.

Chen Ming, “Very interested, space issues, open to data sharing”



**Persona 2**

**Very interested, space issues, open to data sharing**

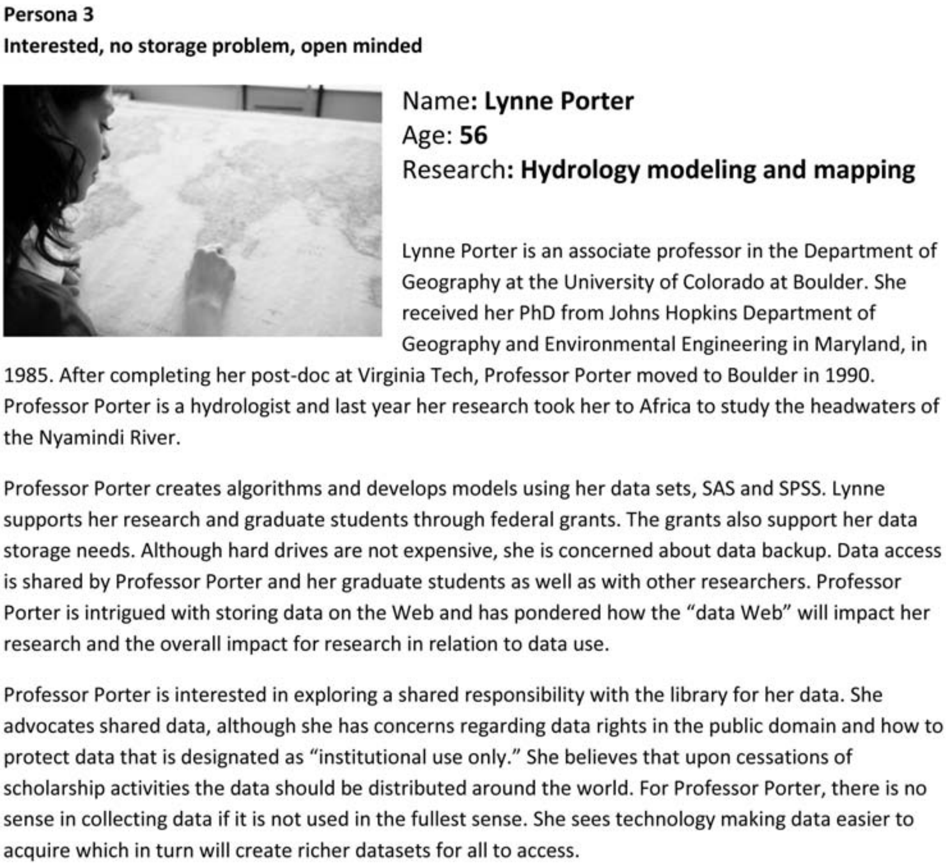
Name: **Chen Ming**
Age: **28**
Research: **Systems ecology; Non-point source pollution**

Chen Ming is a PhD candidate in physical geography. He has a passion for teaching that began at MIT, where he obtained his BS and MS in the first class to graduate in the new 5-year Bachelors to Masters environmental engineering program. During his fourth and fifth years he was the primary teaching assistant for the core ecological engineering courses. Ming’s specialty is hydrology and systems ecology. He was interested in continuing his studies in the Department of Geography in order to include the study of human-induced changes to ecological systems. His area of research is non-point source pollution generation, focusing on mercury.

Ming’s research produces ASCII files, imagery (.jpg files), and simulations (MATLAB files and movie files). For his analysis, he manipulates some of his data using spreadsheets. His data are all electronic. Ming stores his data on the small network of computers in his advisor’s lab. He also uses Google Docs to store and share spreadsheets with his students and collaborators. The lab network is maintained by another graduate student. The computers are backed-up, but occasionally there are technical problems, and support for resolving technical issues can be problematic. Space is never a problem for the smaller ASCII files and spreadsheets, but he is always running into storage space problems for his larger imagery and simulation files—the core of his research data.

Ming is very interested in a library-run data repository. He sees it as a solution to his space problems and as a means to more easily share data with colleagues. He would want the interface to be simple and the system to be robust and to support a variety of file types and access modes. His data is not all open access—some of it may be patentable or not able to be shared at all levels, so he would need the ability to control who has access to the data.

Lynne Porter, “Interested, no storage problem, open minded”



**Persona 3**
**Interested, no storage problem, open minded**

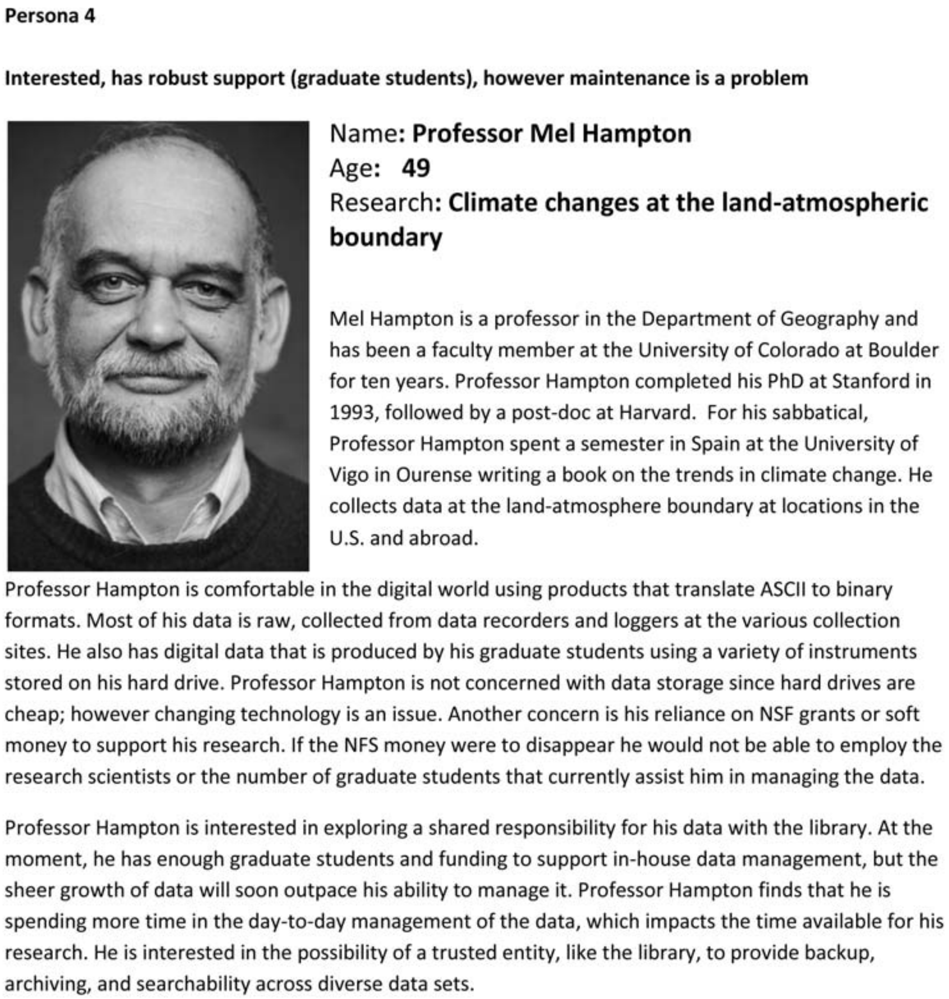
Name**: Lynne Porter**
Age: **56**
Research**: Hydrology modeling and mapping**

Lynne Porter is an associate professor in the Department of Geography at the University of Colorado at Boulder. She received her PhD from Johns Hopkins Department of Geography and Environmental Engineering in Maryland, in 1985. After completing her post-doc at Virginia Tech, Professor Porter moved to Boulder in 1990. Professor Porter is a hydrologist and last year her research took her to Africa to study the headwaters of the Nyamindi River.

Professor Porter creates algorithms and develops models using her data sets, SAS and SPSS. Lynne supports her research and graduate students through federal grants. The grants also support her data storage needs. Although hard drives are not expensive, she is concerned about data backup. Data access is shared by Professor Porter and her graduate students as well as with other researchers. Professor Porter is intrigued with storing data on the Web and has pondered how the “data Web” will impact her research and the overall impact for research in relation to data use.

Professor Porter is interested in exploring a shared responsibility with the library for her data. She advocates shared data, although she has concerns regarding data rights in the public domain and how to protect data that is designated as “institutional use only.” She believes that upon cessations of scholarship activities the data should be distributed around the world. For Professor Porter, there is no sense in collecting data if it is not used in the fullest sense. She sees technology making data easier to acquire which in turn will create richer datasets for all to access.

Professor Mel Hampton, "Interested, has robust support (graduate students), however maintenance is a problem"



**Persona 4**

**Interested, has robust support (graduate students), however maintenance is a problem**

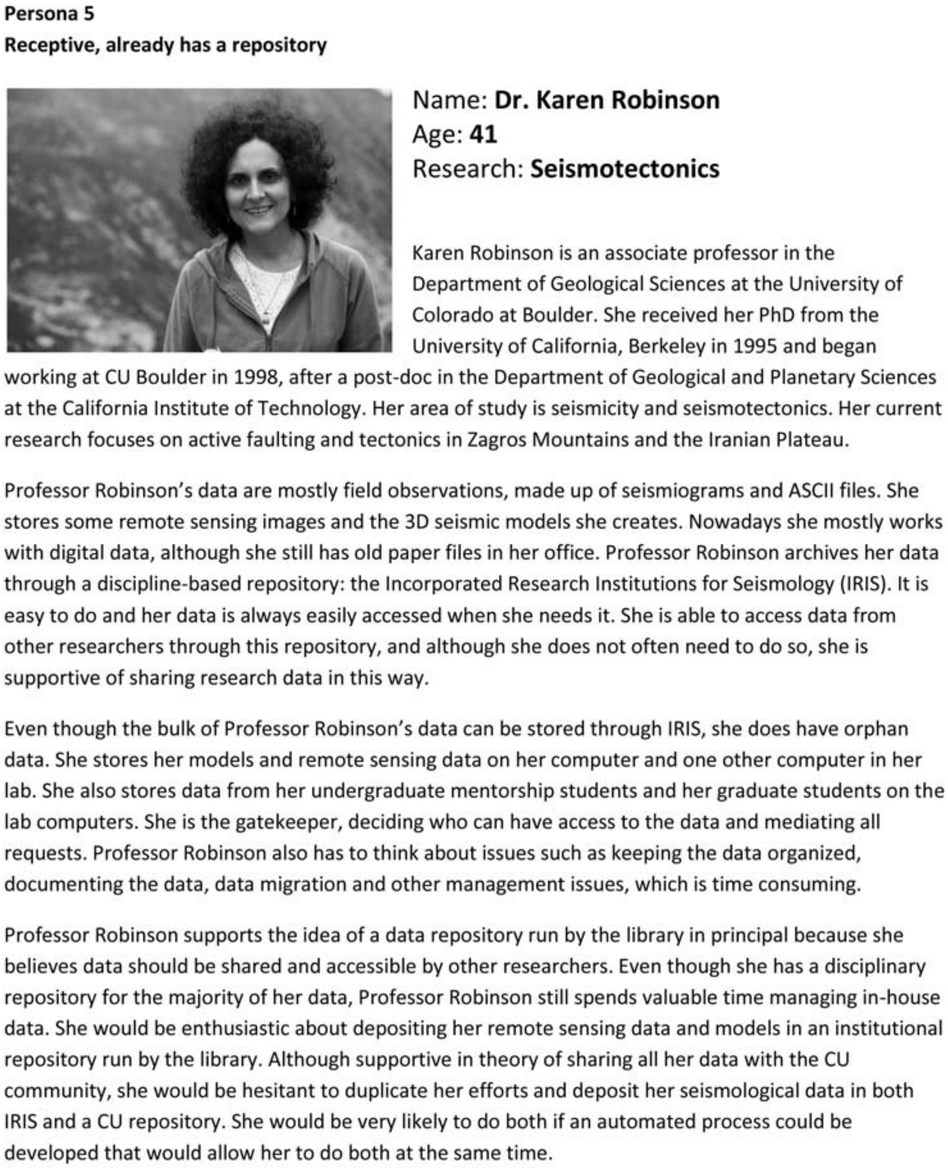
Name**: Professor Mel Hampton**
Age: **49**
Research**: Climate changes at the land-atmospheric boundary**

Mel Hampton is a professor in the Department of Geography and has been a faculty member at the University of Colorado at Boulder for ten years. Professor Hampton completed his PhD at Stanford in 1993, followed by a post-doc at Harvard. For his sabbatical, Professor Hampton spent a semester in Spain at the University of Vigo in Ourense writing a book on the trends in climate change. He collects data at the land-atmosphere boundary at locations in the U.S. and abroad.

Professor Hampton is comfortable in the digital world using products that translate ASCII to binary formats. Most of his data is raw, collected from data recorders and loggers at the various collection sites. He also has digital data that is produced by his graduate students using a variety of instruments stored on his hard drive. Professor Hampton is not concerned with data storage since hard drives are cheap; however changing technology is an issue. Another concern is his reliance on NSF grants or soft money to support his research. If the NFS money were to disappear he would not be able to employ the research scientists or the number of graduate students that currently assist him in managing the data.

Professor Hampton is interested in exploring a shared responsibility for his data with the library. At the moment, he has enough graduate students and funding to support in-house data management, but the sheer growth of data will soon outpace his ability to manage it. Professor Hampton finds that he is spending more time in the day-to-day management of the data, which impacts the time available for his research. He is interested in the possibility of a trusted entity, like the library, to provide backup, archiving, and searchability across diverse data sets.

Dr. Karen Robinson, “Receptive, already has a repository”



**Persona 5**
**Receptive, already has a repository**

Name: **Dr. Karen Robinson**
Age: **41**
Research: **Seismotectonics**

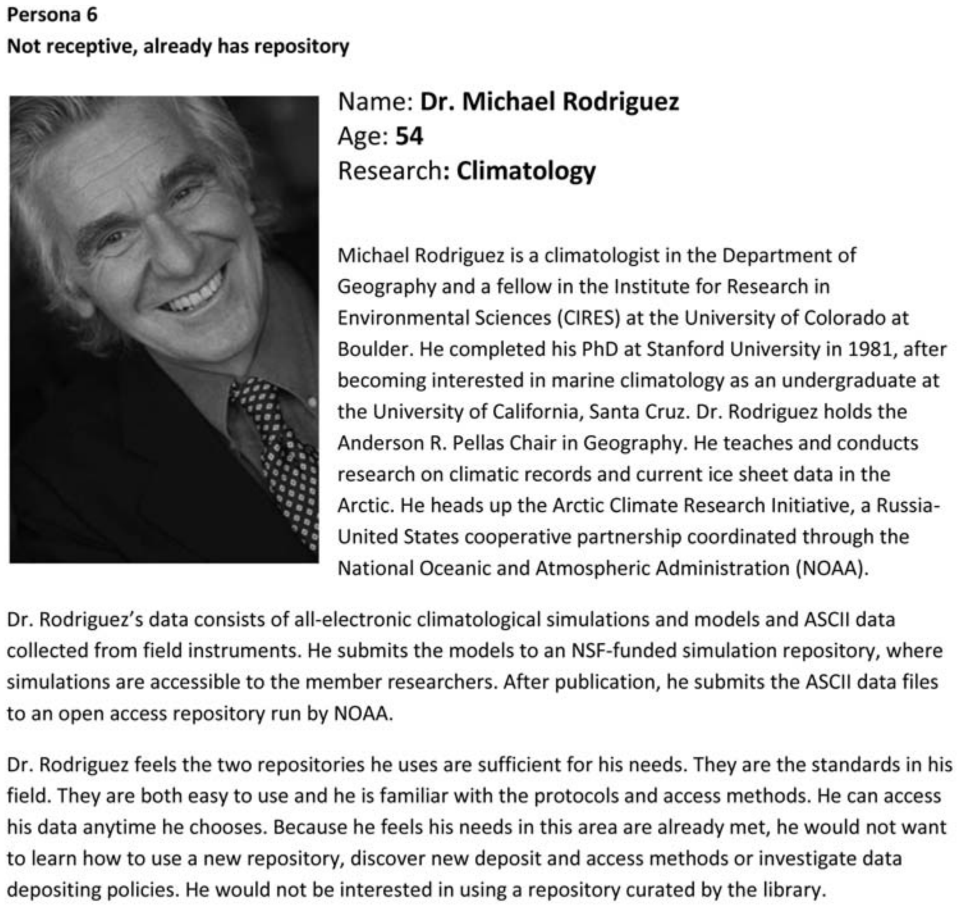
Karen Robinson is an associate professor in the Department of Geological Sciences at the University of Colorado at Boulder. She received her PhD from the University of California, Berkeley in 1995 and began working at CU Boulder in 1998, after a post-doc in the Department of Geological and Planetary Sciences at the California Institute of Technology. Her area of study is seismicity and seismotectonics. Her current research focuses on active faulting and tectonics in Zagros Mountains and the Iranian Plateau.

Professor Robinson’s data are mostly field observations, made up of seismiograms and ASCII files. She stores some remote sensing images and the 3D seismic models she creates. Nowadays she mostly works with digital data, although she still has old paper files in her office. Professor Robinson archives her data through a discipline-based repository: the Incorporated Research Institutions for Seismology (IRIS). It is easy to do and her data is always easily accessed when she needs it. She is able to access data from other researchers through this repository, and although she does not often need to do so, she is supportive of sharing research data in this way.

Even though the bulk of Professor Robinson’s data can be stored through IRIS, she does have orphan data. She stores her models and remote sensing data on her computer and one other computer in her lab. She also stores data from her undergraduate mentorship students and her graduate students on the lab computers. She is the gatekeeper, deciding who can have access to the data and mediating all requests. Professor Robinson also has to think about issues such as keeping the data organized, documenting the data, data migration and other management issues, which is time consuming.

Professor Robinson supports the idea of a data repository run by the library in principal because she believes data should be shared and accessible by other researchers. Even though she has a disciplinary repository for the majority of her data, Professor Robinson still spends valuable time managing in-house data. She would be enthusiastic about depositing her remote sensing data and models in an institutional repository run by the library. Although supportive in theory of sharing all her data with the CU community, she would be hesitant to duplicate her efforts and deposit her seismological data in both IRIS and a CU repository. She would be very likely to do both if an automated process could be developed that would allow her to do both at the same time.

Dr. Michael Rodriguez, “Not receptive, already has repository”



**Persona 6**
**Not receptive, already has repository**

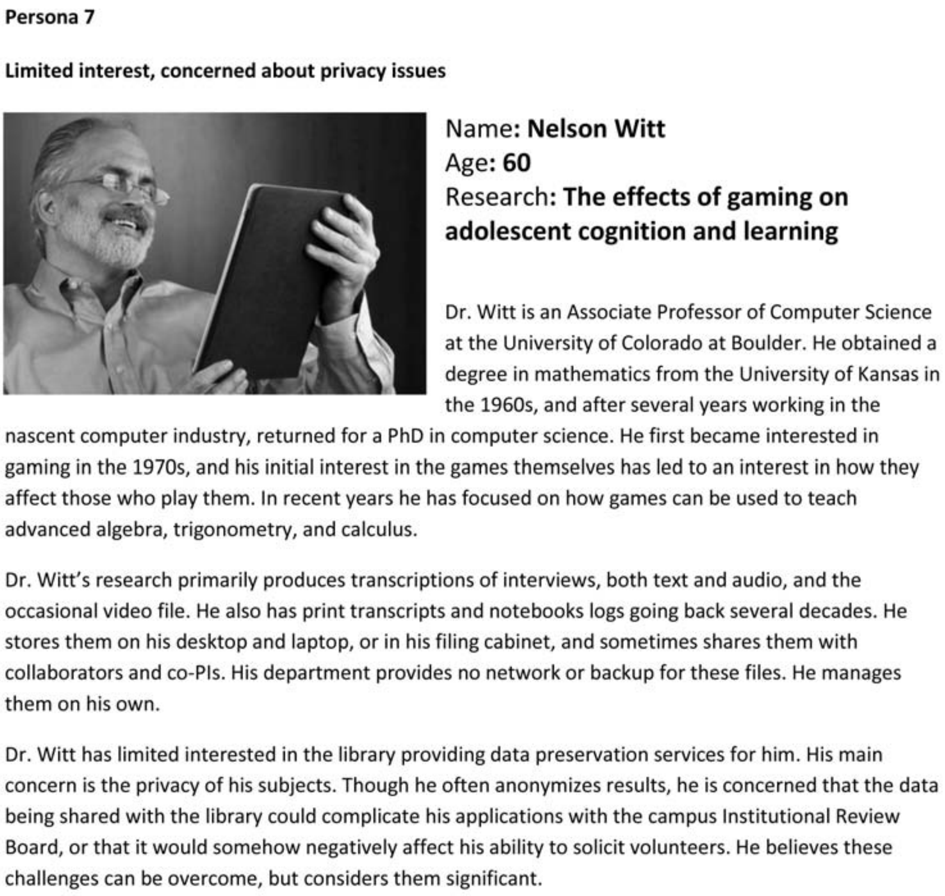
Name: **Dr. Michael Rodriguez**
Age: **54**
Research**: Climatology**

Michael Rodriguez is a climatologist in the Department of Geography and a fellow in the Institute for Research in Environmental Sciences (CIRES) at the University of Colorado at Boulder. He completed his PhD at Stanford University in 1981, after becoming interested in marine climatology as an undergraduate at the University of California, Santa Cruz. Dr. Rodriguez holds the Anderson R. Pellas Chair in Geography. He teaches and conducts research on climatic records and current ice sheet data in the Arctic. He heads up the Arctic Climate Research Initiative, a Russia-United States cooperative partnership coordinated through the National Oceanic and Atmospheric Administration (NOAA).

Dr. Rodriguez’s data consists of all-electronic climatological simulations and models and ASCII data collected from field instruments. He submits the models to an NSF-funded simulation repository, where simulations are accessible to the member researchers. After publication, he submits the ASCII data files to an open access repository run by NOAA.

Dr. Rodriguez feels the two repositories he uses are sufficient for his needs. They are the standards in his field. They are both easy to use and he is familiar with the protocols and access methods. He can access his data anytime he chooses. Because he feels his needs in this area are already met, he would not want to learn how to use a new repository, discover new deposit and access methods or investigate data depositing policies. He would not be interested in using a repository curated by the library.

Nelson Witt, “Limited interest, concerned about privacy issues”



**Persona 7**

**Limited interest, concerned about privacy issues**

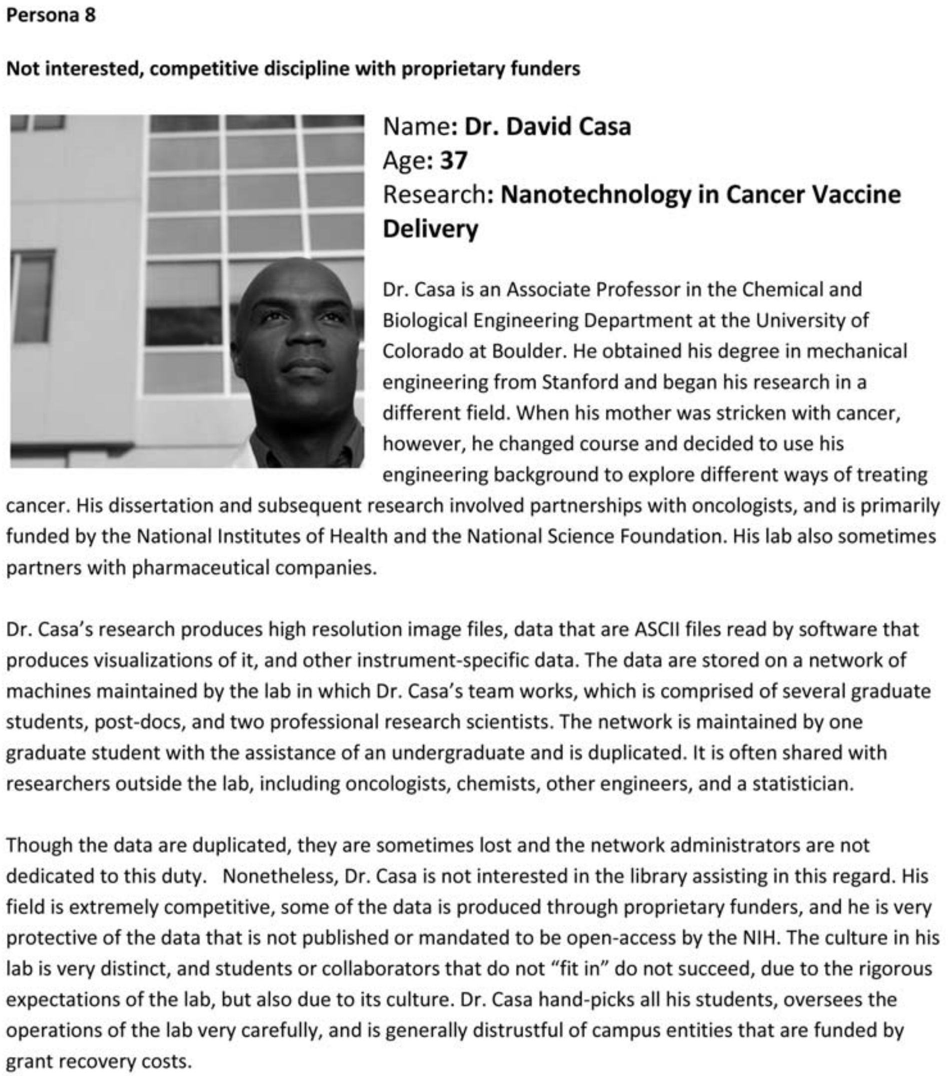
Name: **Nelson Witt**
Age: **60**
Research: **The effects of gaming on adolescent cognition and learning**

Dr. Witt is an Associate Professor of Computer Science at the University of Colorado at Boulder. He obtained a degree in mathematics from the University of Kansas in the 1960s, and after several years working in the nascent computer industry, returned for a PhD in computer science. He first became interested in gaming in the 1970s, and his initial interest in the games themselves has led to an interest in how they affect those who play them. In recent years he has focused on how games can be used to teach advanced algebra, trigonometry, and calculus.

Dr. Witt’s research primarily produces transcriptions of interviews, both text and audio, and the occasional video file. He also has print transcripts and notebooks logs going back several decades. He stores them on his desktop and laptop, or in his filing cabinet, and sometimes shares them with collaborators and co-PIs. His department provides no network or backup for these files. He manages them on his own.

Dr. Witt has limited interested in the library providing data preservation services for him. His main concern is the privacy of his subjects. Though he often anonymizes results, he is concerned that the data being shared with the library could complicate his applications with the campus Institutional Review Board, or that it would somehow negatively affect his ability to solicit volunteers. He believes these challenges can be overcome, but considers them significant.

Dr. David Casa, “Not interested, competitive discipline with proprietary funders”



**Persona 8**

**Not interested, competitive discipline with proprietary funders**

Name**: Dr. David Casa**
Age**: 37**
Research**: Nanotechnology in Cancer Vaccine Delivery**

Dr. Casa is an Associate Professor in the Chemical and Biological Engineering Department at the University of Colorado at Boulder. He obtained his degree in mechanical engineering from Stanford and began his research in a different field. When his mother was stricken with cancer, however, he changed course and decided to use his engineering background to explore different ways of treating cancer. His dissertation and subsequent research involved partnerships with oncologists, and is primarily funded by the National Institutes of Health and the National Science Foundation. His lab also sometimes partners with pharmaceutical companies.

Dr. Casa’s research produces high resolution image files, data that are ASCII files read by software that produces visualizations of it, and other instrument-specific data. The data are stored on a network of machines maintained by the lab in which Dr. Casa’s team works, which is comprised of several graduate students, post-docs, and two professional research scientists. The network is maintained by one graduate student with the assistance of an undergraduate and is duplicated. It is often shared with researchers outside the lab, including oncologists, chemists, other engineers, and a statistician.

Though the data are duplicated, they are sometimes lost and the network administrators are not dedicated to this duty. Nonetheless, Dr. Casa is not interested in the library assisting in this regard. His field is extremely competitive, some of the data is produced through proprietary funders, and he is very protective of the data that is not published or mandated to be open-access by the NIH. The culture in his lab is very distinct, and students or collaborators that do not “fit in” do not succeed, due to the rigorous expectations of the lab, but also due to its culture. Dr. Casa hand-picks all his students, oversees the operations of the lab very carefully, and is generally distrustful of campus entities that are funded by grant recovery costs.

## Discussion

Analysis of the personas shows that many factors affect the receptivity to library involvement in scientific data curation. The authors identified both positive and negative correlations to receptivity.

Positive correlations to receptivity:

- Close proximity to data curation activities. Researchers who were unaware of their labs' curation activities had less inclination toward library involvement. Those researchers who shared some responsibility for this task, however, saw library involvement as a benefit in assisting them in a task that distracted them from their science. Chen Ming, Mel Hampton, and Judy McDannell are examples of how proximity to the task correlates to receptivity. In some cases, the interviewees represented by these personas were graduate assistants who had inadvertently assumed this responsibility for the research group. In other cases they were information technology professionals who believe in the adage "lots of copies keeps things safe."
- Lack of existing curation support. Researchers who had no disciplinary repository and little departmental or lab support or experience in administering networks that allow for automated backup and retrieval systems were more likely to be interested in library assistance. Judy McDannell and Nelson Witt are personas who embody this issue.
- A personal ideology disposed toward sharing. Regardless of all other factors, some personas were receptive to library involvement simply because they viewed data sharing as part of their mission and social obligation as scientists. Lynne Porter is most notable among these.
- Earth sciences research. The interviews demonstrated that researchers in earth sciences, from geologists to environmental engineers, share a disciplinary culture that is more likely to lend itself to partnering not only with librarians, but other labs and researchers, in sharing and providing access to data. Though these interviews did not specifically address access, but focused on curation and preservation of data, access was invariably discussed in most interviews. Lynne Porter, Karen Robinson, Chen Ming, and Mel Hampton are examples of this disciplinary culture. The personas involved in earth sciences, as opposed to applied, life, and physical sciences, tended to be more receptive.[29]

Negative correlations to receptivity:

- Research involves proprietary data. Researchers in disparate fields create and work with proprietary data or funding agencies that require non-disclosure agreements for all or part of the research or data. These researchers are understandably reluctant or entirely unable to participate in a library-run repository. Dr. David Casa represents these researchers.
- Inability to share data in ethnographic research or research using human subjects. Privacy of human research subjects is paramount. These researchers have similar issues to the proprietary research discussed above. Dr. Nelson Witt represents their concerns.

- Extremely competitive field or disciplinary culture that discourages outside involvement. The persona of Dr. David Casa also represents the non-interested researcher on this issue. However, Chen Ming represents a researcher who is more open to exploring the possibility of a library-run repository.
- Existing repository. Personas representing researchers who had a robust existing data repository, often disciplinary-based, were less likely to be receptive to a library-run repository. These researchers, embodied in the persona of Michael Rodriguez, were satisfied with their existing repository and reluctant to learn new data deposit and access policies without the need to do so. (This factor was not primary, however; some researchers who have repositories were still interested, based on other grounds.)
- All researchers expressed a strong reluctance to participate in a repository that was designed in a manner that did not fit their needs and, therefore, would require extra work on their part. Engaging the research faculty about their needs, then, is a critical first step toward repository design.

The use of the personas offers a methodology for identifying receptive researchers that can in turn further the communications for developing a user-centric repository.

The authors had hypothesized a positive correlation between the number of years involved in research with a more traditional mode of research and data sharing, i.e., less receptivity to a library-run repository. However, no reliable correlation was found. The PhD students interviewed were all interested in a library data repository, but many other researchers, including long-time faculty, were as well.

The authors identified possible issues with the methodology used in this study. Participants were not selected at random, which could have introduced bias regarding the researchers' receptivity to the library's role in data curation. However, the results did not suggest that this was the case. Also, the interview questions did not extensively explore the role of the library in data curation; however, at this stage, the authors were primarily interested in receptivity and not necessarily the details of how a researcher might use a repository. Additional studies regarding repository designs and users' needs are necessary.

Ultimately, this study suggests that librarians target researchers similar to the personas of Chen Ming, Judy McDannell, and, possibly Mel Hampton in order to develop data curation partnerships. Karen Robinson and Lynne Porter would be appropriate secondary personas to approach. Personas such as Nelson Witt, Michael Rodriguez, and Dr. David Casa are not likely to be receptive to library involvement in these endeavors anytime soon.

## Conclusion

The research described in this paper furthers the discussion of data curation in libraries by applying a new approach: the use of personas. Models previously described in the literature have been tactical models such as traditional needs assessments or profiles of datasets in order to plan data curation activities. This research instead can be considered a strategic, rather than tactical, approach to understanding not only data, but the scientists and the disciplinary, institutional, and perhaps even departmental cultures in which those profiles and personas work. By offering data curation services to researchers who share traits with the personas, this article suggests, libraries can expedite the partnerships necessary to begin fostering this new realm of science library service.

**By offering data curation services to researchers who share traits with the personas, this article suggests, libraries can expedite the partnerships necessary to begin fostering this new realm of science library service.**

Personas are by definition generalizations. They are created to give a systems designer a concrete user to keep in mind and create empathy, but not to provide specific data regarding small design choices. Different institutions, then, can utilize the personas created in this study to begin or continue strategies for partnering with researchers in data curation initiatives. There are Karen Robinsons, Nelson Witts, and Mel Hamptons at universities across the nation. By developing a stronger sense of the researchers' needs, existing practices, and, most important, receptivity to a library role in data curation, libraries can truly begin providing a new form of library service in a new world of scientific investigation.

*Kathryn Lage is Assistant Professor, Map Librarian, and Acting Faculty Director at the Jerry Crail Johnson Earth Sciences & Map Library, University of Colorado Boulder; email: Katie.Lage@Colorado.edu. Barbara Losoff is Assistant Professor and Associate Faculty Director, Science Library, University of Colorado Boulder; email: Barbara.Losoff@Colorado.edu. Jack Maness is Assistant Professor and Faculty Director, Gemmill Library of Engineering, Mathematics, and Physics, University of Colorado Boulder; email: Jack.Maness@Colorado.edu*

# Appendix

Barbara Losoff, Assistant Professor, Associate Faculty Director, Science Library, Norlin Library
Jack Maness, Assistant Professor, Faculty Director, Engineering Library
Katie Lage, Assistant Professor, Map Librarian, Jerry Crail Johnson Earth Sciences & Map Library
184 UCB
University of Colorado, Boulder
Boulder, CO 80309

Date:

To:

You are invited to participate in a short interview (15-30 minutes) regarding scientific data creation and use at the University of Colorado, Boulder. Barbara Losoff, Associate Faculty Director, Science Library, Jack Maness, Head of the Engineering Library, and I are conducting an organizational data inventory in an attempt to gain an understanding of data production, use, storage and access. This data inventory is motivated by both CU Boulder's proposed institutional repository and the NSF mandates requiring grant recipients to archive and provide access to data.

From the NSF :

- "All science & engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected & preserved" (NSF 2007).
- "The new types of organization envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise…" (NSF Cyberinfrastructure Grants 2008).
- "University-based research libraries and research librarians are positioned to make significant contributions in this area, where standard mechanisms for access and maintenance of scientific digital data may be derived from existing library standards developed for print material." (NSF Cyberinfrastructure Vision for the 21st Century, 2007).

We hope you will consider meeting with us to conduct an interview. Your contribution will inform the Libraries about data on this campus, offer insights for designing CU's institutional repository, and help define the role for the Libraries (if any) regarding data archiving, storage, and access.

## Confidentiality:

If you participate in the survey, your responses will be held in strictest confidence. No identifying links between responses and the individual responding will be retained.

Combined data only will be reported.

Thank you for helping us with this important project.

Sincerely,

Barb Losoff
Barbara.losoff@colorado.edu
303-492-1859

Jack Maness
jack.maness@colorado.edu
303-492-4545

Katie Lage
katie.lage@colorado.edu
303-735-4917

Name____________________
Date_____________________
Department________________
Status_____________________

Briefly describe your research.

How long have you been conducting this type of research?

Can you tell us a little about what sort of data your research produces?

How is the data stored and accessed after it is produced?

Who has access to this data?

Does your department/lab have procedures in place for the preservation of researchers' data in the event they leave the university or pass away?

Is storage space problematic?

Would it be of interest to you for those responsibilities to be transferred to an entity within the university, such as the Libraries?

Please rate on scale of 1-5 your receptivity to this question, 1 = least interested, 5= very interested.

## Notes

1. Daisy Abbott, "What is Digital Curation?" Digital Curation Centre, (2008) http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation (accessed 24 June 2011).
2. Ibid.
3. Tony Hey and Jessie Hey, "E-science and Its Implications for the Library Community," *Library Hi Tech* 24, 4 (2006): 515.
4. National Science Foundation, *Cyberinfrastructure Vision for 21st Century Discovery*, NSF 07-28, http://www.arl.org/bm~doc/ci_vision_march07.pdf (accessed 24 June 2011).
5. Lucy Nowell, "Global Research Library (GRL) 2020 Position Paper" (2007), http://www.grl2020.net/uploads/position_papers/Lucy%20Nowell.pdf (accessed 24 June 2011).
6. National Science Foundation, *Sustainable Digital Data Preservation and Access Network Partners (DataNet)*, http://www.grants.gov/search/search.do;jsessionid=XVTQLvcK0GfcttG1gCVpGny7TV4Xd4Stxg9hyzL6BhhnmRknxtqp!482146507?oppId=45614&mode=VIEW (accessed 24 June 2011).
7. Anna Gold, "Cyberinfrastructure, Data, and Libraries," Parts 1 & 2, *D-Lib Magazine* 13, 9/10 (2007), http://www.dlib.org/dlib/september07/gold/09gold-pt1.html (accessed 24 June 2011) and http://www.dlib.org/dlib/september07/gold/09gold-pt2.html (accessed 24 June 2011); Liz Lyon, "Dealing with Data: Roles, Rights, Responsibilities and Relationships," Consultancy Report, UKOLN, University of Bath (2007), http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx (accessed 5 July 2011); Research Information Network-Consortium of Research Libraries **(RIN-CURL),** "*Researchers' Use of Academic Libraries and Their Services,*" (2007): 1-70, http://www.rin.ac.uk/files/libraries-report-2007.pdf (accessed 24 June 2011).
8. Joyce L. Ogburn, "The Imperative for Data Curation," *portal: Libraries and the Academy* 10, 2 (2010): 241 (accessed 24 June 2011).
9. Tracy Gabridge, "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data," *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* 265 (2009): 15, http://www.arl.org/bm~doc/rli-265-gabridge.pdf (accessed 24 June 2011).
10. Gold, "Cyberinfrastructure," Part 2.
11. D. Scott Brandt, "Librarians as Partners in E-Research: Purdue University Libraries Promote Collaboration," *C&RL News* 68, 6 (2007): 365, http://crln.acrl.org/content/68/6/365.full.pdf (accessed 24 June 2011).
12. Michael Witt, "Institutional Repositories and Research Data Curation in a Distributed Environment," *Library Trends* 57, 2 (2009): 197-198.
13. Tyler O. Walters, "Data Curation Program Development in U.S. Universities: the Georgia Institute of Technology Example," *International Journal of Digital Curation* 4, 3 (2009): 83, 91
14. Gail Steinhart and John Saylor, "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library," (2008), http://ecommons.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf (accessed 24 June 2011).
15. John Hopkins University, "Sheridan Libraries Awarded $20 Million NSF Grant: Funds to be Used for Data Curation," October 1, 2009 http://www.library.jhu.edu/about/news/releases/pressrel09/nsfgrant.html (accessed 24 June 2011).
16. Gold, "Cyberinfrastructure," Part 1.
17. Jennifer Haas and Sharon Murphy, "E-Science and Libraries: Finding the Right Path," *Issues in Science and Technology Librarianship,* (2009), http://www.istl.org/09-spring/viewpoint1.html (accessed 24 June 2011).
18. Association of Research Libraries, "The Research Library's Role in Digital Repository Services: Final Report of the ARL Digital Repository Issues Task Force" (2009), http://www.arl.org/bm~doc/repository-services-report.pdf (accessed 24 June 2011).

19. Karla Hahn, “Introduction: Positioning Liaison Librarians for the 21st Century,” *Research Library Issues: A Bimonthly Report from ARL, CNi, and SPARC*, 265 (2009): 1 http://www.arl.org/bm~doc/rli-265-hahn.pdf (accessed 24 June 2011).
20. Michael Witt, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin, “Constructing Data Curation Profiles,” *The International Journal of Digital Curation* 4, 3 (2009): 93, 96.
21. Alan Cooper, *The Inmates are Running the Asylum* (Indianapolis, IN: Sams, 2004), 123-124John Pruitt, and Tamara Adlin, *The Persona Lifecycle: Keeping People in Mind throughout Product Design,* (San Francisco: Morgan Kaufman, 2006), 744.
22. Frank Long, “Real or Imaginary: The Effectiveness of Using Personas in Product Design,” *Irish Ergonomic Review: Proceedings of the Irish Ergonomics Society Annual Conference* (2009), http://www.frontend.com/products-digital-devices/real-or-imaginary-the-effectiveness-of-using-personas-in-product-design.html (accessed 24 June 2011).
23. Tomasz Miaskiewicz, Tamara Sumner, and Kenneth A. Kozar, “A Latent Semantic Analysis Methodology for the identification and Creation of Personas,” *Proceedings of ACM CHI 2008 Conference on Human Factors in Computing Systems* (2008): 1501-1510.
24. See, for example, http://www.steptwo.com.au/papers/kmc_macquariecasestudy/index.html, http://ecommons.cornell.edu/bitstream/1813/8302/2/cul_personas_final2.pdf, http://kwhitenton.com/libraries_personas/ (accessed 24 June 2011).
25. Jack M.Maness, Tomasz Miaskiewicz, and Tamara Sumner, “Using Personas to Understand the Needs and Goals of Institutional Repository Users,” *D-Lib Magazine* 14, 9/10 (2008), http://www.dlib.org/dlib/september08/maness/09maness.html (accessed 24 June 2011).
26. National Science Foundation, *Cyberinfrastructure Vision for 21st Century Discovery*.
27. National Science Foundation, *Sustainable Digital Data Preservation*.
28. Steve Mudler and Ziv Yaar, *The User Is Always Right: A Practical Guide to Creating and Using Personas for the Web* (Berkeley, CA: New Riders, 2007): 36.
29. These findings correspond with other research. Earth science disciplines are data driven and have a long established mission to “understand the Earth as a system and meet societal needs.” Ann Q. Gates et al, “Towards Secure Cyberinfrastructure for Sharing Border Information,” *Departmental Technical Reports (CS)*, Paper 151, (2006): 6, http://digitalcommons.utep.edu/cs_techrep/151 (accessed 24 June 2011). Jean Bernard Minster, from the University of California-San Diego distinguished professor of geophysics and chair of the International Council for Science’s World Data System Scientific Committee (ICSU WDSSC), spoke about data sharing in the earth sciences, “there is a long history and a culture of data sharing, mostly because the science itself depends critically on such sharing” Jean Bernard Minster, “Scientific Data Sharing Project,” (2010), http://scientificdatasharing.com/earth-sciences/interview-with-jean-bernard-minster/ (accessed 24 June 2011).