This paper begins with a concept: establishing a shared datasets initiative at the University of North Carolina at Chapel Hill’s School of Information and Library Science (SILS). The initiative under proposal would attempt to introduce a collection of digital datasets, as a shared learning object, into courses across the school’s curriculum to augment learning on digital data topics. The program would have four primary components: program oversight, infrastructure, resources, and services. This feasibility study was designed to ascertain how and whether a shared datasets initiative should be designed for the school. Interviews and focus groups were conducted with SILS faculty and students; in addition, several common reading program administrators and data experts were interviewed. Based on these data, specific recommendations are made for the implementation of a shared datasets initiative at SILS. These recommendations address decisions to be made about technology infrastructure, dataset selection, and the importance of visible success.

Headings:

- Information science—Study and teaching
- Library science—Study and teaching
- Digital libraries
- Data libraries
- Computer-supported collaborative learning
- Curriculum enrichment
THE SHARED DATASETS INITIATIVE: ON THE FEASIBILITY AND DESIRABILITY OF ESTABLISHING COMMON DIGITAL DATASETS FOR USE IN AN ILS PROGRAM

by
Christopher J. Weeg

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 Library Science.

Chapel Hill, North Carolina
July 2012

Approved by

_______________________________________
Barbara M. Wildemuth
Introduction

We can easily recognize in the meaning of phrases like “succès de scandale” and “fairweather fan” the power that simple shared human attention holds for directing social dynamics. Belief in the value of people’s attention is a premise that underlies, and is confirmed by, an enormous advertising industry, which generates nearly $200 billion of ad revenue annually in the U.S. (MAGNAGLOBAL, 2012).

Shared attention can, for a time, be directed toward practically anything at all, from an obscure Olympic sport that few people notice except every four years, to a scandalous new online video, to a humanitarian relief effort following some natural disaster like a tidal wave, or an earthquake, or a wildfire.

This paper describes a project that hinges on harnessing the collective attention of a relatively small group of people: the several hundred scholars, scientists, and students who make up the School of Information and Library Science (SILS) at the University of North Carolina (UNC) at Chapel Hill. It prompts this highly localized community to imagine how they might use the asset of their shared attention to help direct the school’s approach to a changing information landscape that is increasingly digitized and increasingly rich in heterogeneous data flowing from sensors and satellites, from hospitals and highways, from Facebook and Fox News. The project asks this group of individuals to set aside a place in the school’s collective consciousness for a few exemplar datasets, so that the school’s
constituents might have a common base of digital objects from which to work: a small group of datasets with which everyone can engage and that is directed to everyone’s attention — to be described, manipulated, analyzed, and shared in the classroom, in the library, in the hallway, online, and in conversation.

The working title of the proposed program is the Shared Datasets Initiative, or SHADi (pronounced “shady”) for short. The concept is to establish a small, unique, and as yet undefined collection of electronic datasets, specially curated and managed to meet the learning objectives of SILS constituents, whether those objectives have been set out in an official way by the school (for example, through published degree requirements), in a semi-official way (perhaps expressed in course syllabi or demonstrated by student research projects), or they are held latently by the school’s faculty and students.

The concept behind SHADi is borrowed from an idea long familiar to the library world — the idea of the book club. In a book club, a group of readers elect to each read the same book at the same time, so that they can all gather and discuss their shared reading experience. Few of us want the entirety of our reading to be selected by committee. But many people recognize the value of allotting some time to read what everyone else in a group is reading, because this commonality can be a prerequisite for the kind of discussion that goes beyond generalities to delve into specifics. Indeed, in the education domain this principle of everyone-reads-the-same-material is played out as a matter of daily business in nearly every classroom at every level from kindergarten to college seminar all across the country. Some
universities today are taking this concept and amplifying it, by annually assigning a shared reading selection to an entire student body. With all the academic diversity on display at universities, programs like these are meant to help unite a campus around an experience that is both shared and academically focused. This idea of the university common book program is the immediate inspiration for a school-wide shared datasets initiative at SILS. The vision for the initiative at this preliminary stage is that it can provide a shared object that is both vehicle and fodder for community building, and that the program can begin to close a gap in the school’s curriculum where education and training on digital data is concerned. The focus of this research project is to talk to SILS faculty and students in order to answer two questions. The first asks in an abstract way to what degree a program like SHADi is warranted at SILS. The second asks for input on how a program like SHADi, were it to be initiated, ought to be designed for greatest usefulness to SILS constituents.

It should be pointed out that the SHADi concept is essentially an idiosyncratic brainchild of the researcher. In effect, this research project is an attempt to reconcile that brainchild with the real world. The researcher devised the concept, which has its own origins and justifications that require explanation. Then, after receiving positive general feedback, he launched into this research project, which investigates the potential for establishing a program at SILS based on the concept. This paper, then, will deal with two fundamentals — the concept behind SHADi, and the possibility of establishing an active SHADi program at SILS, from the perspective of SILS faculty and students.
Background

This chapter is made up of four sections. The first briefly covers some of the research projects and the courses at SILS that are interested in digital data. The second provides a short overview of a phenomenon known as Big Data and discusses how it relates to SHADi. The third section describes the components of the proposed SHADi program. The fourth discusses parallels between the proposed initiative and common book programs.

Data at SILS

“Data” is a major buzzword these days — whether it is Big Data, microdata, metadata, personal data, data security, data analytics, etc. — anytime the topic is digital information. This is as true at SILS as it is anywhere. Data-oriented projects or groups with which SILS faculty have ties include at least the Data Intensive Cyber Environments Center,1 the iRODS data grids group,2 the Renaissance Computing Institute,3 the DataNet Federation Consortium,4 the Digging Into Data challenge,5 the Sustainable Archives and Leveraging Technologies group,6 the Metadata Research

1 http://dice.unc.edu/
2 https://www.irods.org/
3 http://www.renci.org/
4 http://datafed.org/
5 http://www.diggingintodata.org/
6 http://salt.unc.edu/
Center,\textsuperscript{7} the Odum Institute (with its installation of Harvard’s Dataverse technology),\textsuperscript{8} the Cyberinfrastructure for Billions of Electronic Records project,\textsuperscript{9} the Digital Curation Curriculum program,\textsuperscript{10} ibiblio,\textsuperscript{11} the Laboratory of Applied Informatics Research,\textsuperscript{12} and the North Carolina Disease Event Tracking and Epidemiologic Collection Tool.\textsuperscript{13}

On paper these various groups and projects create a sizable footprint at the school. But at present there are few highly visible avenues for most students to get involved with them. Recently a number of special topics courses have been offered that focus on data-related issues, several of which are taught by faculty members who are involved with these research groups. The courses include a NoSQL database course, an iRODS course, and a data management course. Yet relatively few students enroll in the courses or petition to take independent study courses under the tutelage of these SILS research groups. Why this might be is unclear. It is possible that almost no students want to engage in digital data-oriented learning and research, although this seems unlikely at a top information and library science (ILS) program like SILS. Perhaps students are waiting for invitations from faculty and administrators that are not forthcoming. It could be to some degree an issue at the school with enrollment in irregular courses, such as special topics and independent study courses. Maybe it is not obvious to students what relationship might exist

\textsuperscript{7} http://ils.unc.edu/mrc/
\textsuperscript{8} http://www.irss.unc.edu/
\textsuperscript{9} http://ci-ber.blogspot.com/
\textsuperscript{10} http://www.ils.unc.edu/digccurr/
\textsuperscript{11} http://www.ibiblio.org/
\textsuperscript{12} http://lair.unc.edu/
\textsuperscript{13} http://www.ncdetect.org/
between the faculty-led courses in which students can enroll and the faculty-led research that goes on outside the classroom. Or perhaps students simply do not have, do not know they have, or do not know whether they have the technical or research skills to make a contribution to the various SILS-affiliated data research projects. Whatever the case, these groups and projects seem to create a rather smaller footprint on the SILS student body than on the school’s faculty. The SHADi program aims to make getting involved in these activities easier for students by laying a foundation for increased data awareness, data technical skills, and data literacy school-wide.

SILS is considering, but does not currently offer an academic program focused on digital data, as some of its peers and neighbors do, including Northwestern University (Master of Science in Predictive Analytics),\(^{14}\) the University of Illinois at Urbana-Champaign (Master of Science with specialization in Data Curation),\(^{15}\) Syracuse University (Certificate of Advanced Study in Data Science),\(^{16}\) and nearby North Carolina State University (Master of Science in Analytics).\(^{17}\) SILS does offer a certificate in Digital Curation, but the program description is geared more toward an archival perspective than toward a data management or analytics perspective.\(^{18}\) It emphasizes “the ability to plan, manage and implement practices that ensure the long-term integrity and use of resources that are created in digital form,” and two of the program’s five course requirements

\(^{14}\) http://www.scs.northwestern.edu/grad/mspa/
\(^{15}\) http://www.lis.illinois.edu/academics/programs/ms/data_curation
\(^{16}\) http://ischool.syr.edu/academics/graduate/datascience/index.aspx
\(^{17}\) http://analytics.ncsu.edu/?page_id=1799
\(^{18}\) http://sils.unc.edu/programs/certificates/digital_curation
cover archives and preservation, whereas there is no statistics or data analytics requirement. Likewise, there are some SILS courses that deal with specific relevant aspects of data structures and technology, but they range across the SILS core curriculum and across the electives that feed a variety of official and de facto specialty tracks (archives, digital libraries, web technology, cataloging) without any overarching interconnection that would tie the topics they cover into a logical sequence or relationship. These courses include but are not necessarily limited to Organization of Information (INLS 520), the databases sequence (INLS 523, 623, and 723), Programming (INLS 560), Understanding Information Technology for Managing Digital Collections (INLS 465), Research Methods (INLS 780), Information Visualization (INLS 541), Metadata Architecture and Applications (INLS 720), and the iRODS course, Policy-Based Data Management (INLS 624). Even with all the faculty’s work on data, and the various courses available, there is no officially (or even unofficially) recognized or sanctioned academic track to provide guidance in this area. The average SILS master’s student would need to be able to articulate early on the intent to focus on data, and would need to carefully tailor a course of study and practical experience, in order to prepare herself for a leadership position in the increasingly data-hungry business, government, academic, or library worlds.

“Big Data” Buzz

Of all the data conversations in which the information professions are currently engrossed, the buzz around so-called Big Data is probably the most
prominent. The news has spread from technical and academic circles to the popular press, with headlines like:

- At Davos, Discussions of a Global Data Deluge (Bilton, 2012)
- The Search For Analysts To Make Sense Of “Big Data” (Noguchi, 2011)
- Cutting Through the Big Data Hype (Monahan & Griffin, 2012)
- Big Data, Fast & Slow (Stokes, 2011)

The term Big Data, while it sometimes earns initial capitals, is widely agreed to have no known definition. How big is Big, everyone asks rhetorically. A common description, among the kinds of firms that sell Big Data products and services, revolves around the inconvenience or the outright inability of users to store or crunch data without the product or service that is being marketed: “Big data refers to data sets that are too large to be hosted in traditional relational databases and are inefficient to analyze using non-distributed applications,” according to Amazon (Amazon Web Services, 2012, see Overview section), which rents storage space and computing capacity for a fee:

Suppose you host a popular e-commerce website. In order to understand your customers better, you want to analyze your Apache web logs to discover how people are finding your site. You’d especially like to determine which of your online ad campaigns are most successful in driving traffic to your online store.

The web service logs, however, are too large to import into a MySql database, and they are not in a relational format. You need another way to analyze them. (see Getting Started section)

This characterization underscores two attributes commonly associated with Big Data: there is the question of big-ness, and there is the issue of data structure.
Data has the potential to be “big” along one or more different axes. For instance you can have a relatively small number of very large digital objects; or you might have an enormous quantity of tiny digital objects; or maybe some permutation of the two. Either way, a large reservoir of data can be processed more quickly by multiple processors working in tandem than by a single computer working alone, which means Big Data processing responds well to technological economies of scale: it is the domain of large technology-rich firms (who can afford to run parallel processing on large computer clusters) like Amazon, Google, Netflix, Facebook, and LinkedIn; or of powerful research institutions such as the military, NASA, NOAA; or of select well-funded groups at select well-funded universities.

To be sure, some SILS graduates may wind up habitually crunching data as big as the data crunched by these major players. However, the average SILS graduate probably will not, partially because SILS is only beginning to teach the skills and tools required to do so. But in any case big-ness is relative. Any digital data work that ties up a computer processor (for instance, on a desktop computer) for more than a moment or two enters a realm in which data-size is salient to the user: it is big enough that its size becomes part of the issue when opening, saving, storing, updating, moving, or analyzing it. Almost every SILS graduate will encounter this problem (or already has), and in this way the “relatively big” is a problem shared by all.

As far as data’s structure is concerned, the heterogeneity of web-based transaction logs is just one version of more or less unstructured information. There are untold numbers of other possible examples. Scientific sensor data, video and audio bit streams, a list of files on a laptop computer, the free text of a monograph or a journal article, the
tags used in a family’s digital photo album, the pixels in a single digital image — these and many other forms of digitized information could be considered relatively unstructured digital data. But none of them are necessarily Big Data. Small objects, or digital objects in small quantities, are as capable of embodying the spectrum of unstructured to highly structured as is a large amount of digital data. The two dimensions exist independent of one another. Still, Big Data is often what gets talked about and, as a highly visible symbol for issues of both data size and of data structure, it is probably worthwhile to talk about it, especially in terms of ILS curricula.

In addition to the technical dimensions of Big Data, there is an ethical dimension that also ought to be of interest to students and instructors of the information professions, and that deserves to be unpacked. In a provocative interrogation of the subject, boyd and Crawford (2012) point out several sticky intellectual and ethical dilemmas raised by the “era of Big Data.” What assumptions, they ask, about knowledge and research are embedded in the quantifying approaches Big Data seem to favor, and what might these “models of intelligibility” do to the research enterprise if we allow them to “crystallize into new orthodoxies” (p. 666)? Research, information, knowledge — these are always loci of power, and there are fundamental questions about who wields it when it comes to Big Data.

Much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data.

But who gets access? For what purposes? In what contexts? And with what constraints? While the explosion of research using data sets from social media sources would suggest that access is straightforward, it is anything but. [...] 

Some companies restrict access to their data entirely; others sell the privilege of access for a fee; and others offer small data sets to university-based researchers. This produces considerable unevenness in the system: those
with money — or those inside the company — can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access. (boyd and Crawford, 2012, p. 673-4)

Big Data, then, has the potential to reinforce and further entrench digital divides. If private firms, for whatever complex of reasons, guard the gates to proprietary stores of massive data, and if they hold the keys to the proprietary data computation engines, then they (and their bottom lines) get to dictate what becomes known and what counts as knowable at the confluence of research and high technology. SILS students and graduates will be far better prepared to work with and around Big Data if they have the opportunity to investigate these issues. And even those who may never actively become involved in Big Data research themselves, as consumers of research and as ambassadors for information access, they need to be able to navigate these murky waters. Understanding the principles of up-to-date data technologies and the terms of the political debates surrounding them (or even the semantic debates around buzzwords like Big Data) should be somewhere on the list of learning objectives for schools like SILS.

This stance is a basic tenet of the initial SHADi program proposal. A basic objective of the initiative, in its preliminary articulation, is to cultivate learning opportunities for SILS students in areas related to data technology — including topics like statistics, programming languages, file management, and data literacy. The proposed strategy for SHADi is to intervene in the SILS community with an initiative that combines data-oriented resources, infrastructure, and services
designed to raise awareness of data technologies and to promote data literacy among the school’s students, both in and out of the classroom.

**Program Components**

There are four main components to the proposed SHADi program: resources, infrastructure, services, and general program oversight (see Figure 1). These are sketched out here in a general way, in order to provide a working picture of the core of the proposal. More details related to these components, informed by the findings of this study, are addressed in later sections of this paper.

![Figure 1. The four components of the proposed SHADi program.](image)

**Program Oversight.** In this context, the concept of program oversight refers to governing the SHADi program, i.e., the instrumentation of leadership on constitutional issues like mission and core values, which lay the groundwork for the other three components. The oversight work will most likely be executed by a team or a committee formed for the purpose. In the beginning, the oversight component
will necessarily be focused on program architecture, coming to decisions on key issues that will strongly influence resource selection, infrastructure design, and direction of services. These issues will often address “should” questions. Should SHADi exist? To what degree should SHADi attempt to integrate itself into courses (and which courses)? What kind of funding or other institutional support should be sought? What should the resource collection policy be? The other three components of the SHADi program could then be seen as primary interest areas to be governed by steering decisions made by the program oversight team. Management of these interest areas would likely be delegated to what will be referred to here as SHADi personnel. Once SHADi were to be implemented, it is likely that another important duty of program oversight would be program evaluation.

**Resources.** The proposed SHADi resources are first and foremost datasets. What exactly a dataset *is* remains an open question, but for this project the term refers, if vaguely, to some collection of recorded information, the focus here being on digital data. This could be a table or set of tables from a relational database. It could be a corpus of text or of bibliographic records; scientific data of some sort; or network data. It might be a data-stream from some interesting application programming interface. It could be a collection of digital photographs, videos, or sound recordings. (Examples are helpful to illustrate the possibilities, but at this stage are not meant to proscribe them.) The idea is that a dataset be a digital object (or a collection of digital objects) that coheres in one way or another into a “set” of data. The possibilities are nearly endless; the SHADi oversight team will need to
determine what datasets (and how many of them) should be put into the system for the greatest chance of success of the initiative.

**Infrastructure.** The program’s infrastructure will need to encompass some sort of technology by which the datasets in the collection can be shared among the school’s constituents (primarily students and faculty). Again, the selection of the infrastructure technology will need to be determined by the SHADi oversight team. It could be as simple as hosting a collection on the SILS website, from where the resources could be downloaded by anyone. This is a less interesting and less dynamic possibility than some, because the options for interaction with the datasets and with the collection as a whole, and the options for communication between program participants, are quite bare. Considerably more intriguing options might include UNC’s Sakai e-learning environment or some version of SILS’s own Lifetime Library technology. Also included under the infrastructure umbrella would be whatever personnel can be found to maintain the initiative’s technology and to run its services, which of course is one way of saying that the program will realistically require some degree of funding.

**Services.** Finally, the services involved will cover three main areas. The first coincides somewhat with program infrastructure: the program personnel who run the system will do some level of administrative and maintenance work. This work is likely to include setting up and populating with datasets whatever the SHADi environment ends up being; possibly creating and updating user profiles; working out whatever technical kinks inevitably appear; and generally keeping an eye on the
technical aspects of the program. The second type of service would be service to faculty members who would like to incorporate the SHADi datasets into the courses they teach. While some instructors may have the ability and the inclination to take care of this themselves, others may not see how or why to insert the datasets into their instruction practices. As a major branch of the foreseen user population of SHADi, the SILS faculty deserve explicit consideration for user services. The third form of service is activities and events planning and staffing. The SHADi program ought to have a footprint outside of the classroom (guest speakers, technology tutorials and troubleshooting, research consultations, poster sessions — there are many possibilities), and it bears acknowledging that administration of these activities is a service that must be planned for.

These four components of the SHADi program are proposed to give the initiative a recognizable presence in, across, and outside of SILS courses. In practice, implementation of the program will look different from various perspectives. It is almost certain that some courses simply will not use SHADi resources or services: not every course can gracefully incorporate a digital dataset into its discussions, assignments, exams, or projects; not every instructor cares to use every new tool or collaboration or system that is made available. This is expected. Some courses may make passing reference to it. It is not hard to imagine datasets like Association of Research Libraries (ARL) statistics or ISI Web of Knowledge citation indexes being put into, or referenced from, SHADi. Instructors for courses like Library Assessment (INLS 782) or Resource Selection and Evaluation (INLS 513) might present these to students for demonstration purposes. Other courses may make more involved use of
the initiative. A database, a metadata, a data/text mining or a programming course, for example, could build semester-long projects out of using and manipulating one or more SHADi datasets. Outside of the school’s regular courses, students could use the datasets for a variety of enrichment purposes and for supervised independent study and master’s paper/project courses.

**Lessons Learned from Common Book Programs**

While it is easy to see SHADi as fundamentally a digital library, the concept for the program can perhaps best be understood as an extension from, or an analogy with, university common book programs. As was mentioned above, these programs are centered around prompting very large groups of students to all read the same book. The programs often go by one of several related names, including one book programs, common reads programs, summer reading programs, book-in-common programs, or any number of variations on this theme. There is no official definition, and every common book program is unique, but what these programs generally do is, first nominate a book that they encourage or require some university constituency to read, and then provide those constituents with a structured opportunity to gather and address their common reading experience. Very often the constituency is a university’s incoming first-year student cohort, so typically the reading is meant to happen over the summer prior to students arriving on campus for the first time. When students do arrive, there is often a kind of introductory period lasting several days, during which they get situated in their dorm rooms, attend various orientation, advising, and social events — and also participate in
discussion groups, led by university faculty and staff, that are focused on the common book selection. It is also typical that a convocation speech or keynote address will be given by the book’s author, or by someone else who can address issues the book raises.

There is apparently little research that has been done on these kinds of programs. Much of the literature that does exist is advice for the would-be practitioner or anecdotal accounts by people who have run a common book program (often it is both). The University of South Carolina houses a National Resource Center for the First-Year Experience and Students in Transition, which provides resources for summer reading programs, including a listserv, a database of institutions that have tried common book programs, and lists of books that have been used. The institute publishes a monograph series covering all kinds of issues pertaining to first-year programs, number 44 of which is Common Reading Programs (Laufgraben, 2006). The book is a practical guide for planning to initiate or to improve a common book program, and its arguments don’t necessarily seem to be evidence- or research-based. Relevant citations are mostly to the websites or newsletters of universities with common book programs.

With a lack of published research on the topic, and without any official governing documents, the entity of the common book program can best be described as a phenomenon, or a trend, or a concept. Any institution can obviously take or leave any component that it cares to, or reinvent the idea to fit local

---

19 http://sc.edu/fye/
20 http://sc.edu/fye/resources/fyr/srp/index.html
conditions. There is no federated list of all participating institutions, nor a published survey, so it is not currently possible to put a number on “how popular” common book programs are, much less track the concept’s growth over time. Anecdotally, the database of participating institutions mentioned above lists some 487 instances of common book programs (with many institutions appearing multiple times, once for each year on record), with no reason to believe that this list is exhaustive.

Since at least 1999, UNC has been running a common book program of its own. The university ran into significant controversy over its 2002 selection, *Approaching the Qur’án*, which garnered media coverage nationwide (the controversy, not the book), as students, parents, politicians, pundits, and religious groups complained. Three students filed a lawsuit alleging the program infringed UNC students’ “religious free exercise” by forcing them “to study Islam against their will,” according to a July 2002 press release from the Family Policy Network, a Christian group that aided in the suit (Family Policy Network, 2002). Since then the common book program, known as the Carolina Summer Reading Program, has merely encouraged UNC students to read its selections, rather than required them to do so. The program’s 2012 summer reading book is Nicholas Carr’s 2008 bestseller, *The Shallows: What the Internet Is Doing to Our Brains*.

There are some features of the common book program concept, as it is popularly practiced, that do not translate to the intentions of the SHADi program. Among these are its focus on undergraduate students and a semi-remedial emphasis

21 http://summerreading.web.unc.edu/
on modeling basic college skills: a common book program inserts an academic-type consideration into the social and transactional business of moving from high school to college, and “sends the message — early in a student’s transition — that reading and discourse are expected in college” (Laufgraben, 2006, p. 3). There are probably different basic assumptions at play when graduate students are the audience for a program, and so SHADi is less concerned with pump-priming for academic attitudes, so to speak, and more concerned with enriching the experiences of highly able students already dedicated to scholarship. Another difference is the common book program’s typical timeline, which usually spans the summer with a brief set of interactions limited to the beginning of the fall semester. It is a goal of the SHADi program to be an all-year presence, integrated into regular courses. This goal is probably the most difficult facet of the program to realize, because it requires broad consensus across the school, and it asks many instructors to tailor their course content and the collection content toward agreement with one another. But the benefits envisioned by a highly integrated SHADi program are far superior to the fruits of a non-integrated version, which would be unlikely to garner much notice or use. Whether the benefits of an integrated SHADi program outweigh the costs of implementing it is a question for SILS to decide.

While there are clear differences between common book programs and SHADi, the programs potentially share several features. Laufgraben (2006) highlights six building blocks of the common book program concept: reading selection, activities, academic focus, skill development, shared leadership, and theme (pp. 12-15). Each of these can easily be applied to the concept behind SHADi,
and their enumeration could be helpful for setting out program design parameters. The reading selection of course modifies to datasets selection. Academic focus is self-explanatory, and activities have already been discussed as a component of SHADi services. Skill development, as has already been mentioned, incorporates a range of learning objectives from statistics understanding to technical know-how. The concepts of shared leadership and theme bear some expansion. In the first case Laufgraben emphasizes the need for collaboration: every group that will be involved with the outcome of planning should be allowed to provide input on program design and resource selection, probably through involvement in planning committees. For the SHADi program, SILS faculty and students are likely to be the major parties involved, which is why these populations were selected for interviews and focus groups. In the second case a theme can be used both to direct planning decisions and to help lubricate conversations with program participants:

> Before or after a book or readings are selected, the common reading program may take on a theme around which resource materials and activities are planned. [...] An overarching theme for the selection can help tie perspectives together and help students understand the rationale behind a choice. These themes are then interpreted and explored from different disciplinary perspectives. (Laufgraben, 2006, p. 13)

Following this lead, establishing a theme for the SHADi program might help either to create a rationale for selecting datasets, or once a collection of datasets is chosen, to cultivate interest among constituents and partners within and outside of SILS by providing an evocative framework for communicating about the initiative; or it could be used for both of these purposes.
Methods

The data collection component of this research project involved speaking with topic experts and with SILS constituents, to get their ideas on whether and how a program like SHADi could and should be initiated at the school. The data were collected in two ways (semi-structured interviews and focus groups) from four distinct populations (SILS faculty, SILS students, common book program administrators, and data experts, i.e., librarians and other professionals whose jobs focus on working with data and with university researchers interested in data). The number of study participants of each type, involved in each type of data collection, is summarized in Table 1.

Table 1. Number of subjects and number of research sessions by population type and data collection method.

<table>
<thead>
<tr>
<th>Population</th>
<th>Total subjects</th>
<th>Interview subjects</th>
<th>Focus group subjects</th>
<th>Focus group sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SILS faculty</td>
<td>9</td>
<td>7</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>SILS students</td>
<td>10</td>
<td></td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Common bk. admins.</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data experts</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>12</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: One faculty member participated in both an interview and a focus group.

Twelve interviews were conducted: seven faculty members were interviewed; two common book program administrators; and three data experts.
There were three focus group sessions: two student focus groups; and one faculty focus group. One of the student focus groups had three participants, while the other had seven participants. The faculty focus group had three participants, one of whom was also interviewed. Thus the total number of research subjects was 24 people.

Each population type was asked interview questions attuned to that group’s context (see Appendix A for the interview guides used). Common book administrators were asked questions about their specific program’s history and administration. Data experts were asked about the data services and resources offered at their institution. SILS faculty were asked about course design and about historical, current, and potential future use of datasets in their courses and in general at the school. Data experts and SILS faculty were both asked open-ended questions regarding their impressions about, or advice for, a shared datasets initiative proposal.

The focus group questions were likewise tailored to subject populations (see Appendix B for the focus group discussion guides used). Faculty were, again, asked about course design and about the potential for course adoption of the SHADi datasets. Students were asked about their experiences with and attitudes toward various co-curricular activities, about translating common book programs into a common datasets program, and about how to design the SHADi program to be appealing and useful to students.

All SILS interview and focus group sessions were conducted on the UNC-Chapel Hill campus; data expert and common book administrator interviews were
conducted, in each case, on the campus of the university that employed the participant. In addition to UNC these included Duke University and North Carolina State University.

The interview and focus group sessions were conducted by the researcher, and recorded using a digital audio device. The recordings were then transcribed by the researcher. Subsequently these transcripts were hand-coded, using a descriptive coding system devised inductively in response to the transcript content. In practice, there are basically two parallel coding systems, one for the data experts and SILS groups, and one for the common book program administrators. This scenario arose naturally because the content of the participant comments ended up being split along these lines. The coding system for the common book administrators contains 10 codes; the data expert and SILS groups coding system contains 21 codes. These codes, then, became a way to describe on an aggregate level what issues participants spoke about, which groups spoke about which issues, and what they had to say about those issues. The list of codes, along with brief descriptions, is included as Appendix C, and a numerical analysis of their occurrence and co-occurrence is included in Appendix D.

The numbers in Table 1 (above) demonstrate that the data collection was approximately equally focused on interview and on focus group participants, and that SILS constituents played a greater role in the research than topic experts did. The topic experts were included largely for informational purposes, to give the research project access to experience-based knowledge regarding issues basic to the
administration of a common thing program, and to data-oriented research at the university level. The data collected from these subjects serve largely a background or advisory role.

The most critical problems this project tackles are what datasets the SHADi project should include, and how the initiative should be designed in order to best serve the students and faculty of SILS. For this reason, SILS constituents make up most of the research subject sample. The facts that SILS constituents are the intended beneficiaries of the SHADi program and that they will be the group to ultimately design and implement the initiative (if it is to see the light of day) prompted this researcher to model the project on a family of qualitative research methods known as action research, which has variants with names like participatory research, participatory action research, and friendship as method. The work of social psychologist Kurt Lewin is often cited as an early progenitor of action research (Berg, 2009; Ferrance, 2000; Levin, 1999). Lewin helped bring overt researcher involvement into the social research equation, for instance in a 1940s study in which he “designed an experiment to test if it was possible to encourage American housewives to use tripe as part of their normal, everyday meals” (Levin, 1999, p. 26). Commentators note that this branch of social research also has historical ties to feminist theory (Gillies & Alldred, 2002; Tillman-Healy, 2006) as well as to political activist theory and practice, including labor and class movements in developing countries (Fals-Borda, 1991; Rahman, 1991); it is about using research to empower institutional change. There is also a significant body of work using action research to look at the classroom: “For the past several decades, the
practice of action research has been a fairly common mode of investigation in education research, especially among researchers interested in classroom teaching practices” (Berg, 2009, p. 246). Ferrance (2000) describes a taxonomy of four levels of classroom-based action research: undertaken by or on behalf of an individual instructor; a collaborative group of instructors; an entire school; or an entire school district. The second and third of these four levels are described more fully in Table 2, since the SHADi program most likely falls somewhere in the spectrum of the middle two categories.

**Table 2. Types of action research; adapted from Ferrance (2000, p. 6).**

<table>
<thead>
<tr>
<th></th>
<th>Collaborative Action Research</th>
<th>School-wide Action Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus</strong></td>
<td>Single classroom or several classrooms with common issue</td>
<td>School issue, problem, or area of collective interest</td>
</tr>
<tr>
<td><strong>Possible support needed</strong></td>
<td>Planning/release time; close link with administrators</td>
<td>School commitment; leadership; communication; external partners</td>
</tr>
<tr>
<td><strong>Potential impact</strong></td>
<td>Curriculum; instruction; assessment; policy</td>
<td>Potential to impact school structure; policy; evaluation of programs</td>
</tr>
<tr>
<td><strong>Side effects</strong></td>
<td>Improved collegiality; formation of partnerships</td>
<td>Improved collegiality, collaboration, and communication; team building; disagreements on process</td>
</tr>
</tbody>
</table>

By approaching data collection as an opportunity for collaboration and conversation, this research project aims to articulate realistic expectations and possibilities relative to the variety of needs and desires that members of the SILS community bring to the concept of a shared datasets initiative — and in this way facilitate the emergence of a live program that will benefit the school.
Findings and Proposal Development

As noted above, this research project is intended to address four primary components of the proposed SHADi program: oversight, resources, infrastructure, and services. To the extent that this paper serves as the first step in a proposal to the SILS community that the school adopt the SHADi program, these four components also represent an initial articulation of what might be called the proposed initiative’s problem profile: if the school wants to go forward with the initiative, what are the key decision points that must be worked out first? This paper argues that there are four fundamental dimensions of action here, corresponding to the four program components. The first question to be answered is who will participate in program oversight. Once this is determined, the oversight team will need to set out a vision and a mission for the program. From there, they and perhaps others will have to answer questions about what resources to select, what infrastructure to utilize, and what services to offer. Once these decisions are made, the school can begin to actualize the program. If SILS in fact elects to establish a SHADi program, there will undoubtedly arise further decision points, in addition to the four that are covered here, that will ultimately need to be addressed. But at minimum a plan governing oversight, resources, infrastructure, and services will probably need to be laid out before any significant action can be taken toward a first instance of a shared datasets initiative.
Oversight

Composition of oversight team. The first problem is this: who will decide what SHADi is for? Who should be on the program oversight team that will design the initiative's architecture? As one faculty member put it, what would be “the lines of authority”? This is a question that common book administrators could help answer because they already have experience with a university program designed to unite students and faculty around a common learning object. In both cases the common book administrators involved in this study describe selection committees as the bodies that more or less govern their institution’s common book program. At UNC-Chapel Hill there are nine committee members, and at North Carolina State University (NC State) the committee “consists of twenty-something representatives from various units across the campus.”

These committees have a great deal of power to steer the programs, selecting readings and planning activities based on their conceptualizations of the purpose of a common book program. But hidden somewhat between the lines in both common book administrators’ comments, there is reference to a level of authority above the selection committee. At UNC-Chapel Hill, the program was recommended in 1999 following a high level, campus-wide “intellectual climate report” by a committee that reported to the university provost. Subsequent to this creation, the program’s decision-making has been managed by a rotating committee of three staff, three students, and three faculty members. At NC State, the common reading program governing committee originated as a council of academic deans, which after a few
years passed the torch to a less highly placed, and a more diverse, group of university members, referred to at this point as the selection committee.

The point here is that in both cases, the originating spark behind the common book program was embodied by a steering committee operating on a fairly high plane of official authority (i.e., deans and provosts) — certainly higher than the authority enjoyed by current selection committees (i.e., faculty, staff, and students) at either institution. A potential lesson here for the SHADi program is that it will likely be necessary to, at least conceptually, separate large-scale governing operations from the management of implementation, such as selecting datasets and planning an activities calendar. (Managing the selected datasets, and staffing the activities would then be yet another level down.) There are steering decisions that must be made about what SHADi’s objectives should be, and these decisions will influence the implementation decisions.

So, how big should the SHADi oversight team be, and who should be on it? In agreement with Laufgraben’s advice in the book Common Reading Programs (2006), these types of programs seem to have found success with committees whose size and shape are logically related to categories that people recognize as having a stake in program outcomes. At UNC the focus is on affiliation: three each of the faculty, staff, and student categories. Discernible in NC State’s less strict attitude toward committee formation (“[people] can invite their way in, it’s not a closed group”) is the program’s inherited origin as a project of academic deans, i.e., the concept that
any participating group (i.e., a college) is welcome to have a representative onboard. (There are twelve semi-independent colleges at NC State.\textsuperscript{1})

As an organization, SILS has a number of built-in categories that might be of use in forming a SHADi oversight team, such as faculty, staff, undergraduate students, and graduate students. The school has four degree programs, Bachelor of Science in Information Science (BSIS), Master of Science in Information Science (MSIS), Master of Science in Library Science (MSLS), and Ph.D. in Information and Library Science, each with a faculty member serving as program director. There are also a number of standing committees (with faculty and students as members) built into the school's bylaws, including an Undergraduate Program Committee, a Master's Program Committee, and a Research and Doctoral Committee. Additionally, there is a provision in the bylaws for the creation of \textit{ad hoc} committees that can be created to consider “issues that require committee deliberation when the issue does not fall within the confines of any existing committee” (School of Information and Library Science, 2010, p. 10). In the current conceptualization of the SHADi program, there are also three program components (other than the oversight component) that it may be useful to consider, i.e., resources, infrastructure, and services. Attempting to combine some of the SILS-oriented categories with the SHADi-oriented categories could be a useful approach to designing the oversight team. There are several ways to achieve this type of combination, but a potentially attractive way to do it could be to create a nine-person team consisting of three faculty members, one member of the SILS information technology staff (SILS IT),

\textsuperscript{1} http://www.ncsu.edu/academics/colleges/
and five students (two undergraduates, one MSIS student, one MSLS student, and one Ph.D. student). With three program components and nine members, each component could be delegated as an interest area to a sub-group of three team members (see Table 3).

**Table 3. A possible structure for the SHADi oversight team.**

<table>
<thead>
<tr>
<th>SHADi interest area components</th>
<th>Resources</th>
<th>Infrastructure</th>
<th>Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oversight team membership</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty</td>
<td>Faculty</td>
<td></td>
<td>Faculty</td>
</tr>
<tr>
<td>BSIS</td>
<td>SILS IT</td>
<td></td>
<td>BSIS</td>
</tr>
<tr>
<td>MSIS or MSLS</td>
<td>Ph.D.</td>
<td></td>
<td>MSIS or MSLS</td>
</tr>
</tbody>
</table>

In this sketch of a possible oversight team composition, each sub-group could be chaired by a faculty member. (Perhaps invitations for the chair positions could be extended to a member of each of the three standing committees mentioned, i.e., the Undergraduate Committee, the Master’s Committee, and the Research and Doctoral Committee.) Because of the technological nature of the initiative, having someone from SILS IT on the oversight team would probably be advisable, and it makes the most sense to put this person in the infrastructure sub-group, where a familiarity with SILS technology would be put to greatest use. The Ph.D. student has been placed in the infrastructure sub-group primarily because this researcher assumes most SILS Ph.D. students who would be candidates for inclusion in the oversight team will tend to have a better understanding of computer technology than most SILS master’s or bachelor’s student candidates. Also, because SHADi resources and services would be targeted at a greater number of undergraduate students and
master’s students, it probably makes sense to reserve spots in those sub-groups for them.

This oversight team could be formed as an ad hoc committee at SILS. Each sub-group would be able to meet semi-independently to investigate and to discuss its own interest area (potentially including greater numbers of students in these types of sub-group meetings). And the oversight team would be able to then convene as a whole in order to negotiate and to finalize SHADi steering decisions.

There are of course other possible oversight team composition designs, but this one serves to illustrate a few design elements that it might be wise to keep in mind when composing the team. One element is to include members from key SHADi constituencies, and to give team members responsibilities that match their knowledge, abilities, and authority. Another is to find a way to use existing SILS structures to get SHADi oversight work accomplished. A third is to consider symmetries within the team, such as dividing the team up into working sub-groups, that could help make the logistics of governing as pain-free as possible.

**Determining program mission.** After an oversight team is formed, the basic question they must answer is, what are the core values of the SHADi program? What is it for? Who is it for? What should its relationship be with SILS courses? Should it be created in the first place?

No participants in this study made comments to the effect that SHADi should not exist. However, focus group participants were self-selected, in the sense that they replied to email listserv announcements. Perhaps only people who thought
sharing datasets at SILS is a good idea responded. And all participants knew that they were speaking with the inventor of the concept, so they may have been unwilling to say they thought it was a bad idea, even if they thought it was a bad idea. Some participants registered skepticism that the program would be able to get off the ground, as this comment (about using shared datasets for teaching) from a faculty member shows:

Faculty_01: That, I think, is a slightly harder sell. Not that it wouldn't be useful, but you know. You know what faculty are like. We have a hard time agreeing on what the content of even our core courses should be sometimes. [...] The idea of creating a shared dataset that could be used as a teaching tool, while I think it would be useful, I think would be difficult to arrive at an agreement on.

Nevertheless, the same person also said about the program: “We absolutely should be doing something like this.”

One descriptive code that was used in coding transcripts is “Mesh w/SILS,” which was attributed to comments pertaining to the question of how well the proposed SHADi program fits with the purpose or goals of SILS. This idea was discussed 33 times in the interviews and focus groups and occurred at least once in each interview and focus group. These comments demonstrate a range of positive perspectives, from how the initiative could help with community building (Student_01: “I think it really bridges IS and LS really well”), to how it serves the school’s current agenda (Faculty_02: “I think it does mesh, I mean, I think it has all the sort of ideals of SILS”), to the ways it coincides with where SILS would like to go in the future (Faculty_03: “The school’s trying to put together this data management
emphasis, and so this is good timing I think”). Regarding the question of whether SHADi should be created, comments like these potentially affirm that at the very least the concept is worth pursuing further, to determine whether the positive potential many people see in the initiative can overcome the likely disagreements on how to implement it.

What do SILS constituents want SHADi to be like? One of the most intriguing types of comment that repeatedly appeared throughout the transcripts was coded as “be significant.” These comments gravitate toward a few topics. One of them comes primarily from students, and it gives voice to frustrations about a desire for practical experience in the classroom, for something that bridges the academic world and the post-graduation world:

Student_02: Seems like it would be really good to have something available that we could actually use to see an example of how the thing, the theory that we're working on is applied. [...] I mean, it would've been nice, if we're talking about exporting things in XML from a database, to watch somebody do that. And then try to do it ourselves. Same for digital preservation, which I took. Be nice to look at some digital data and see how things that we're learning would actually apply. So like if I'm going to apply the OAIS reference model to some data, what does that look like?

Student_03: Yeah. I agree. It's not just one class, but like, it seems like almost every class you could use this in. I don't know! I just feel like we're doing a lot of reading about this stuff, but you're not actually doing any of it. And this seems like a way that you could actually like, do something.

Another comment thread in the “be significant” category has to do with making connections outside of the school, and even outside the university. One faculty member mentioned a huge collection of recently released World Bank data that the World Bank would be delighted to help SILS use. Another floated the idea of
sharing SHADi with other ILS schools in North Carolina. A third mentioned having already gotten leads on partnering organizations from a member of the UNC-Chapel Hill Board of Visitors. And Student_04 made a comment about the potential to benefit from this style of networking, when a faculty member might “have a tie to [some] other institution. And they could have us working on meaningful things.”

It is this concept of working on meaningful things that is perhaps the most profound — possibly transformational — issue that came up during data collection for this study. It gets to the heart of the admonition to “be significant.” It is something that may be buried in the comments of many participants, but that found clear expression in a couple of them. Here a faculty member talks about providing some meaningful learning opportunities by using “real” datasets:

Faculty_04: From a learning point of view, we’re always learning test cases, we’re doing sort of experiments where the outcomes are well known. They’re exercises, they’re not real problems. And it seems to me that people are much more engaged in real problems. And so, having real data, it’s got to be inherently more motivating than something that’s artificial. Now, from the teaching point of view, having the artificial is more predictable, more controllable, so I could see the advantages. But I mean, from the learner’s point of view, who would want to, I think you’d want to deal with stuff that, you know, you could make discoveries. It’s what makes SETI at home, or the Sloan telescope, Skytel telescope, and all these citizen science kinds of things so powerful. People participate and get excited about it because, you know, they feel they’re contributing to, or working with something that, it’s not just an exercise.

Here a student talks about the potential for deep engagement if students were to be on the receiving end of those kinds of learning opportunities:

Student_04: What I immediately thought you were talking about was moving from sort of these fake ideas like, oh, I’m going to [catalog] my CD collection, and then never use this database or whatever I created ever again. To
something that would be more lasting. And so, in my mind that would have to be real data that then people would want to use. And so what I thought of immediately was citizen science. Something that people really could get into and say, I'm making an impact by analyzing this data. Or doing something, or organizing the data in some way that's more useful. [...] I would be attracted to the idea if I felt like, oh, now my SILS projects really mean something.

This concept of being significant, in every one of these cases, speaks eloquently to the idea that students and faculty come to SILS because they want to get involved with real-life problem solving; they want to make contact with the wider world. The SHADi program has the potential to introduce a new avenue for this kind of contact, and this is a value that many study participants discussed as a factor that ought to be worked in to the initiative's design.

Outside of the value SHADi could hold for SILS constituents in the abstract, there are two primary outcomes that the initial program proposal suggests the initiative could or should be used for. One of these is the idea of helping students learn about data technologies and enhancing general data literacy. Several comments made by data experts indicate that these are key areas when it comes to providing data resources and services. Students made many comments, such as the one quoted above, about wanting to learn these technologies hands-on, and about being frustrated by a lack of opportunities to do so. There are fewer faculty comments on this topic, but one faculty member made a point that the National Science Foundation (NSF) now requires every NSF-funded research project to have a data management plan, and that crafting these plans represents an in-demand technology-oriented skill that ILS students really should be learning.
The other primary outcome that the initial SHADi proposal advocates is an increased facilitation of exchange between SILS constituents. As it stands, SILS students and faculty are meaningfully connected to each other probably mostly as a network, e.g., a student who takes four courses in a semester regularly attends class with about 40 to 60 other students, who each attend class with 40 to 60 other students, and so on. If you zoom out, you get a network diagram of the whole school, that changes over time. The proposed idea with SHADi is to establish a second relationship structure that is shaped more like a hub-and-spokes, with the SHADi datasets as the hub, and every SILS constituent arrayed around the wheel rim. There would be at most one degree of separation between any two community members, student or faculty, regardless of whether they share any courses. Further, this arrangement could remain stable for as long as SHADi was to exist. Of course, electronic datasets shared online are very different from living human beings gathered in the same room at the same time: the quality of relationship that SHADi can offer is different than one made in the classroom, but it opens up possibilities not available otherwise. This, though, is an abstract way of describing how SHADi could affect relationships at SILS. In their comments on this subject (and on all subjects, really), students focused on concrete things. They talked about the sorts of co-curricular events they might or might not attend, and whether or not it would violate the Family Educational Rights and Privacy Act (FERPA) to let students access and witness each other’s academic work through a centralized system. One student compared SHADi to a blog, where people could discuss topics or interests held in common, and also to a debate tournament, where students could interact through
challenges. But overall students did not seem to pay much attention to the social implications of enacting a shared datasets initiative. Faculty paid more overt attention, with two faculty members independently mentioning that SHADi could facilitate communication between instructors on course design and the use of datasets as instructional resources.

Taken together, these findings likely indicate that the emphasis on data literacy and technology is important to faculty and especially to students at SILS, but that an emphasis on the social interactions facilitated by a centralized system like the SHADi program is either not particularly valued, at least by students, or maybe is too abstract to be compelling — a takeaway being that it may not be wise to attempt to use the social aspects to drum up interest in the program. It could be that interest in this as a research topic might develop over time as social dynamics are concretely played out, and could perhaps be measured in some way.

Resources

The concept of a shared datasets initiative only begins to take concrete shape as the question is addressed: what are these datasets? How many will there be? How will they be chosen? Data selection was the second-most discussed topic in the interviews and focus groups, just behind the closely related topic of how datasets could be integrated into courses. Indeed, comments on these topics frequently co-occurred. Comments on data selection also co-occurred fairly frequently with comments on issues between faculty and datasets, issues about getting the initiative started, likely problems, and the initiative’s mesh with SILS. These categories of
comments represent a kind of constellation around the basic problem of, how do you select datasets that the faculty as a group will want to start using in their courses (because it won't be easy)?

**Purpose of resources.** As was discussed in the previous section, many SILS constituents are in favor of using data that is “real” instead of “fake.” Here Data_Expert_01 talks about the value of exposing statistics students to a data archive: “They have to have something to generate frequencies and generate regressions. And what better to use than an active dataset, rather than some made up data that comes built into the software tools.” Similar comments came from SILS faculty:

**Faculty_05:** I’ll tell you what I would dearly love for my students. I would dearly love a digital archival collection. I’d really like a real one. [...] I’d really like one that’s like, it comes in and it hasn’t been processed, right? Because we can go over and look at them all [already] processed. Well, whoop-de-do. I would like some collection [...] that you could actually pull stuff out of, and process. You know, go through the archival processing of the digital collection. And you could do that over and over again.

They also came from students:

**Student_05:** Immediately the kind of thing that I started thinking about was our 520, Organization of Information class, where [...] we were tasked with basically creating our own fake collection. And it was all kind of weirdly hypothetical. We still got some things out of it, but how great would it have been if we could've pointed to this, you know, corpus of data or whatever, and been asked, take some small segment of this and develop it into a collection.
Related to this issue, several study participants compared the SHADi datasets to a Text Retrieval Conference (TREC)\(^2\) test collection (Harman & Voorhees, 2006), but there was some disagreement among comments about whether TREC test collections are “real” or not.

**Selecting a theme.** This question of real-ness asks about the qualities SHADi datasets should exhibit, but at some point the problem of which specific datasets to include has to be addressed. As with common book programs, one way of approaching this topic is to identify a theme for resource selection. A possible strategy would be to hitch a ride on “UNC’s [new] two year pan-campus theme, ‘Water In Our World,’ which calls for the University and all of its members and resources to mobilize around a common issue facing our society, which in this case is water” (Barber, 2012). Because the university at large is already set to be attuned to this issue, and because water is relevant to so many social and political contexts beyond the university, a water theme could help facilitate the kinds of meaningful interactions that so many participants voiced a desire for. This was picked up in a student comment:

```
Student_06: A lot of what we learn here can be transferred into any field, or any endeavor. And it might be a way for us to sort of learn how to sell that. In other words, I could go to the biology department and say, hey we've got this water dataset, um, do you guys ever do any stuff having to do with this? Is there a way that we can put this to use for you?
```

Other students and several faculty members made similar comments. But a few faculty members cautioned that it could be problematic to focus on datasets

outside of the domains with which most SILS students and faculty are familiar. One instructor often has students use the popular Internet Movie Database\(^3\) (IMDb) as an in-class dataset because students can capitalize on the deep familiarity most of them already have with the movies. This allows them to navigate the dataset easily, and to remain focused on the learning objectives of the class, rather than causing students to focus on the terms of some unfamiliar domain. On a related note, during a focus group session another faculty member complained about a remorseless lack of domain knowledge in the current Big Data community, which prompted this exchange:

Faculty_06: Right now there’s these generic people [i.e., non-domain experts], and they pull stuff [from datasets]. Have no idea what they’re looking at. And they don’t know how to make sense of it. So it’s this really bizarre quantification where people keep misreading stuff.

Faculty_03: See, in that way, that would argue for us not doing water. Because we don’t have, maybe one chemist in the whole student body.

Faculty_06: Well, you need somebody who understands —

Faculty_03: We need a dataset that we have some sense of understanding.

From both the positive and the negative comments about using a water theme for the SHADi datasets collection, it is apparent that there are advantages and drawbacks to the concept. This will probably be true of any theme that might be chosen, and from a wider perspective it will probably also be true about any dimension along which selection decisions are made. For instance, consider the quantity of datasets to be included. Many faculty members made comments about this. There is essentially a spectrum with two possible extremes: include only one

\(^3\) http://www.imdb.com/
dataset, or include an unlimited number. The fewer datasets you place in the collection, the harder it becomes to get everyone to agree on which one or ones should be selected. The reflex, then, is to add more datasets to the list. But there are problems in this direction of the spectrum as well, such as overtaxing SILS’s server capacity, increasing administrative overhead, and potentially threatening the shared status of the datasets in the program’s collection. Several faculty members saw pedagogical and logistical advantages in the promise that eventually students would start coming to courses already familiar with SHADi datasets from previous courses. There is a risk, then, in humoring the natural push toward adding more and more datasets, which is that it could begin to prevent datasets from overlapping across multiple courses:

Faculty_04: I do think you’re right, though, about [including] multiple databases.

Faculty_03: But not too many, otherwise I don’t gain any of my advantage from having the students having been exposed to it in a class before.

Clearly, as with the theme, there are advantages and drawbacks involved with any decision about the number of datasets to include, or going further, about the number of years or semesters that a collection of datasets should remain in the collection. The key is to find a compromise that is at least acceptable to the vast majority of SILS constituents.

Properties of resources. One way of thinking about these problems, as was alluded to above, is to parse them out into dimensions or properties. Whether the question is about quantity of datasets or about length of time for inclusion, the
relevant property may be something as simple as *degree of restriction*. What should SHADi’s approach be toward restriction? Given the choice, do we tend toward restriction, or tend away from it? Shall we err on the side of fewer datasets and shorter periods? Or shall we err on the side of more datasets and longer periods? These are questions of values that this researcher cannot answer for the SHADi oversight team nor for SILS, but it is a fundamental issue that should probably be addressed by the oversight or resources committee early in the process and documented for future reference.

This idea of identifying properties to aid in decision making could also be useful for selecting specific datasets to include in the collection. A faculty member who specializes in data and datasets repeatedly made comments insisting that the first thing SHADi should do is to identify the formal properties of data that students and faculty would like to address, and then find datasets that fit those criteria:

Faculty_07: I would even go one step back and say, what are the properties of datasets? What are the properties of datasets which I am going to look at, and how am I going to find exemplars with those properties? So I would even list, these are the properties which you are coming to look at. Like one of them might be a large collection. Ok, size of the collection. [...] Formats, is another one. Formats is an important criteria. Usage is another one. How do people use these different type of things? So you look at, these are the ten properties which I am going to [choose], and I am going to find examples of those things.

Several other faculty members, as well as the study's data experts, made comments that seem to rely on similar assumptions, especially in reference to properties like data type or format. There were many comments in which participants either mentioned or compared various categories of data that might be
desirable, such as numerical, spatial, free text or string data, photographs, videos, sound recordings, relational database records, structured versus unstructured data, and so on. A good next step toward specific dataset selection might be for the SILS faculty or a SHADi oversight team to devise a list of favored properties for which exemplars could then be found. Creating that list will ideally probably involve a content analysis of as many course syllabi as possible, in order to figure out what properties of data are topical in the classroom. This is something several faculty members suggested would be needed in order to cultivate faculty interest. It might also be advisable to conduct further formal data collection aimed at faculty (e.g., a survey, interviews, focus groups) or informal investigation among faculty (e.g., discussion at faculty meetings, etc.), to help determine what are properties of data that ought to be covered. Such work is outside the scope of this research project.

**Infrastructure**

There are two basic branches of infrastructure that would need to be worked out for the proposed initiative.

**Technology infrastructure.** The first of these is the underlying digital technology. How should the SHADi program be implemented, in terms of computer hardware and software? Minimally the collection’s datasets would need to be made accessible somehow to SILS constituents. This could potentially be achieved by creating an accessible set of pointers that refer to datasets existing online elsewhere. The solution here could be as simple as publishing a list of hyperlinks. The risk with this approach, of course, is that the links might break for any number
of reasons, in which case the datasets would functionally be lost. Some broken links might be repairable or replaceable, but others might not be. If SILS aims to be able to decide these sorts of access questions for itself, copies of the initiative’s datasets would need to be stored on a computer over which the SHADi oversight team has control. This proposal assumes that the SHADi program would want to guarantee access to its datasets, and so assumes that the initiative would obtain a local copy of each dataset in the collection. In this scenario, although the datasets could conceivably be stored on a computer that a user may sit at in the SILS library, or be passed around on an external hard drive, the obvious choice for delivery is online via the Internet. Participants seemed universally to assume that this would be the method to use.

However, this approach still leaves open several possibilities. This research project has not pursued in-depth technical research, so an overview of these possibilities relies on this researcher’s general understanding of technical systems likely to be available to SILS. The first of these is the SILS website\(^4\). It is certainly conceivable that the SILS webmaster could create a webpage from which users could download datasets or portions of datasets. Another option might be to create a Sakai project page, where again, users could download data and datasets. A third option could be to request some space on the Odum Institute’s Dataverse Network for SHADi datasets, or to nominate datasets that already reside there. A fourth option could involve the iRODS software behind the Lifetime Library. Several faculty

---

\(^4\) http://sil.s.unc.edu/
members and students spoke positively about this possibility, as in this comment from a faculty member:

Faculty_08: I would like to think of the Lifetime Library as the common dataset. [...] Within the Lifetime Library, it’s very easy to share data. So I can set up a collection that’s accessible by all people in a group, like members of SILS, and therefore SILS can control its own content, and what it wants to distribute between themselves. Or an individual can have a personal library within the Lifetime Library, and choose whether or not they want to share its contents with other students. So the challenge is, I’ll phrase it this way. The Lifetime Library exists. It’s sitting there ready to use. The challenge is coming up with the set of data that are of most interest, that would elicit interest from faculty and get people using it.

To what degree other decision makers on the faculty and staff at SILS would agree with this rosy characterization of the potential for symbiosis between the Lifetime Library and the SHADi project is an open question, and it is one that should probably be explored further. In any case, it seems to be a promising indication that there are viable options for the initiative’s selection of technology.

Technology selection is an important question because the technology that SHADi uses will influence the basic shape of the program. One possible program dimension that both faculty and students brought up on several occasions has to do with a concern about how interactive the shared datasets collection would be. Some participants considered being involved in the conversation about dataset selection as a type of interactivity. Others took it further, trying to imagine whether and how students would be able “to contribute stuff into the common pool,” as faculty member Faculty_01 put it. Here a student imagines this as an elective, extracurricular kind of activity:
Student_06: [I]t would be neat if people could contribute related datasets. Maybe not to the core information that you have, but as an auxiliary thing, like, oh I happened to be on the UN website and I found this stuff about water, let me just upload it, put it in a separate folder. But it could be a supplement to what already exists. And then you know maybe, so maybe people begin to take some ownership of it in that way, too.

A faculty member expressed an idea that others also shared, which is that it might be possible to somehow delegate or offer operational responsibility for system upkeep or for dataset processing and description directly to students through their participation in relevant courses:

Faculty_05: I think you could rig something up that would work for some classes. Yeah. So, you could have something where the metadata people would create metadata for it, right? You could have something where the preservation people were going to look at preservation issues.

Here a group of students discusses the issue:

Student_05: You were [talking] about adding to this body of information. And that seems like that might be really interesting. The idea of like, instead of a theme, it’s oriented more around a problem? And the dataset is kind of dynamically constructed by the, the students, or by the group.

INTERVIEWER: Sort of collecting evidence in a courtroom.

Student_05: Exactly. And then presentations. I mean, and it would have to be like a problem, like capital-p, you know?

Student_06: Crime, or —

Student_05: Right. Right. I mean, it would probably be very similar. Water could very well be the problem. Uh. But it might be interesting to do it that way.

Precisely how this arrangement might work, in administrative terms and in terms of students being graded, would be open to debate. But it serves to illustrate
how deeply interactive the program could become, given the proper environment and infrastructure.

The choice of technology infrastructure, then, ought to be informed by a consideration of what amount of interaction is allowable and is desirable. Posting datasets online for download occupies the low end of the spectrum, but this solution probably comes at a low cost. Providing every student with a personal Lifetime Library, with which both the school and students can create sophisticated read and write permissions, currently occupies the high end of the spectrum, but it involves complications. For instance, tailoring read and write permission rules for various populations (e.g., SHADi administrators, course instructors, and students) across multiple digital objects may not be simple. Finding the right spot for SILS along that spectrum is a decision that, again, the SHADi oversight team would need to make.

**Personnel infrastructure.** Besides technology, the other basic branch of SHADi infrastructure is personnel. Because the initiative is proposed for SILS, and because it is fundamentally techno-centric, there are likely to be very important interfacing issues between SHADi and SILS IT. Any portions of the program that were to attempt to utilize SILS IT resources, including its staff, would obviously need to be worked out with them very closely and very early. This research project, however, takes an agnostic view about the extent to which SHADi personnel might overlap with SILS IT staff. One possible way to structure this interaction would be to include a member of the SILS IT staff on the SHADi oversight team, as was noted in the Oversight section of this chapter (see Table 3), in which case SILS IT staff would
be well positioned to provide input on this matter. The aim here, though, is to begin to delineate the kinds of tasks that would need to be done, leaving for later the sketching out of SILS jurisdictional boundaries. One of these kinds of tasks is certainly the administration and maintenance of the digital technology on which SHADi is to be hosted. Several faculty members did mention SILS IT staff by name in this regard. Others suggested that a graduate student could serve as an administrator. One faculty member illustrated the point by naming a Ph.D. student who currently administers twenty iRODS data grids. In any case, the SHADi program would need to procure some sort of technology personnel who would be able to run the technology portion of the program’s infrastructure. Considering that participants who foresaw this need primarily named SILS IT staff and doctoral students in their comments, it seems likely that this category of personnel would need to be a paid position, either through salary, if only by default (i.e., if it became another task that fell to SILS IT), or through some kind of institutional support (i.e., if a Ph.D. or even a master’s student did the work).

In addition to technology personnel, the other major personnel category would be outreach and services personnel. The sorts of services SHADi might offer will be addressed in the next section of this paper, at which point it will be more appropriate to discuss who will perform them. These personnel are mentioned here as a way to acknowledge that the procurement of service work requires SHADi to engage people to do that work. Preparing this type of staffing infrastructure is a mode of planning that would need to be addressed, to some degree, as infrastructure *qua* infrastructure.
Services

In this paper's brief description of proposed SHADi program components, there are three categories of services listed: technology services, services to faculty, and activities services. Technology services were discussed in the previous section; this section will focus on faculty services and activities services.

Faculty services. Services to faculty members is a topic that generated many comments, perhaps not surprisingly, from faculty members. One of the proposed program's major concerns is negotiating the integration of SHADi datasets into SILS courses, which would undoubtedly be a tricky business. It should be stated clearly that this proposal does not advocate requiring or pressuring faculty to adopt the program into their courses. Instead, the question is how to promote the program's use, and to how to design resources, infrastructure, and services from the outset in ways that would attract use by both faculty and students.

One of the difficulties with integrating SHADi datasets into courses is that it would take a significant amount of work. If the idea were to ask faculty to make reference to the SHADi datasets in their lectures and assignments, instructors would need to spend time and energy exploring the datasets, exploring their course syllabi, and then making changes to their teaching materials. This is a theme that many faculty members expressed in comments like these:

Faculty_02: Well, it takes time to change that kind of stuff, right. Because you have to re-do your slides and re-do your examples and everything.

Faculty_01: Its an ease of use problem, fundamentally. One of the common themes in discussions among the faculty about changes to the curriculum, to the program, to policies and whatnot is, what extra burden will it put on us
and the office staff? Right? Not to say that we're shying away from work. But we're all busy, the office is understaffed. It's like, is it more work?

The threat of SHADi either being, or being perceived as, an increase to faculty workload was widely seen to be a possible impediment to the initiative's success.

Faculty members predicted that their peers would be reluctant to adopt new teaching methods, even potentially valuable ones, if their current methods are working. The concept came up in comments like this:

Faculty_06: The problem with using common resources usually for me is, how can I make this work for me? And if you've got something that's kind of functioning, you just stick with it.

And this:

Faculty_04: It's going to be so dependent on the faculty member. And their perspective. And, you know, somebody who's got a course that's working pretty well, it's going to be a bigger kind of investment for them to try to plug this in.

A relevant question then arises: how can SHADi lower the overhead that it would take for instructors to adopt the program? This is where services to faculty would likely come in. In faculty comments, there are basically two approaches that were discussed. One is to have SHADi services personnel approach individual faculty members (presumably this would be faculty members who have exhibited an interest) with specific suggestions already in mind, probably based on analyses of a specific course's content and objectives. Faculty member comments disagree about how desirable this might be, as is shown in this focus group exchange:
INTERVIEWER: But what about the idea of saying, well I've looked at your syllabus, here's what I think would work well in your class?

Faculty_06: I love it when somebody does that. If somebody wants to do it!

Faculty_03: I don't know. I think that's an exceptional view. Most of them [i.e., faculty members], I think they want to control completely what goes on in their classroom.

Faculty_06: But I mean, I still have the control, but if somebody comes up and says here’s some suggestions —

The disagreement seen in these comments is on one hand a simple difference of opinion, but for faculty members who might see outside influence as an invasion of academic autonomy, this difference could be a serious one. Common book programs have struggled with the same issue. One common book administrator makes a point along these lines in this comment:

Com_Bk_Admin_01: It's hard, I'll be honest. Even in, because I know that my counterpart in Academic Affairs [...] tries to be an ally and advocate. But even she is reluctant, as a senior associate dean in the College of Arts and Sciences, to push that. And a lot of it just has a lot to do with — and I understand it and respect it — the academic freedom of faculty to have that classroom, you know, it's their purview. I think what we try to do is educate about, here's this program, here's the value of it, and hopefully your students have already read this book, so why not have a way to incorporate it. I think where you get more buy in though, like I said earlier, is when you have a book on a topic that connects to whatever they’re researching or studying or teaching in that particular semester. That’s the ideal.

In line with this ideal, the other kind of service to faculty that arose in a number of faculty comments is the idea of making it easier for faculty to do the integrating work themselves. This could mean facilitating collaboration among the group to come up with ways to integrate SHADi datasets into individual courses, into clusters of courses, and into the curriculum as a whole. One faculty member
envisioned this as an addition of metadata to SHADi datasets, which could carry information about how other faculty have used a dataset in their courses. Another faculty member made a comment about bringing that kind of information to a social media format:

Faculty_03: A crowd-sourcing solution to that might work, where you have some kind of repository of ideas, of teaching ideas, where individual instructors can [say], this is what I’m thinking about doing. And other people will comment on it, or [post] their own ideas.

One faculty member saw a need for a kind of instructional service, something that would lay out in clear terms how a faculty member could use the SHADi datasets and infrastructure:

Faculty_01: [T]o say, you know, such and such a tool is available for you to use. And here’s what you would have to put into learning to use it, deploying it in your class. Here’s what it would take, here’s the support behind it.

This low-key style of service provision, where it was mentioned, was universally approved of by faculty members, which seems to indicate that it may be wise for SHADi to focus on non-invasive forms of service to faculty. Incidentally, it also coincides with what one common book administrator described as a way to facilitate communication and innovation among diverse groups participating in a common book program. The organizers behind Com_Bk_Admin_02’s program try to act as a conduit to “[let] all the colleges know what each of the other colleges is doing” in relation to common book-oriented activities. It creates an environment where “this college might see what that college did, and try to adapt it. Again, its still fairly dynamic [...] but it still is, uh, kind of united but separate.” This kind of
integrated autonomy might be something SHADi should aim to emulate, in light of
the faculty’s mixed attitudes toward more direct intervention by SHADi personnel in
classroom planning.

A few faculty members went as far as to imagine who might be able to
provide some of the services they would like to see from the SHADi program. Again,
where it came up, everyone seemed to think that, as faculty member Faculty_09 put
it, “a graduate assistant or two” would be the way for SILS to get people in these
roles. Here is an exchange from the faculty focus group session:

Faculty_03: Do you think if we had a grad assistant assigned to be like
instructor support for use of the datasets? That would help, wouldn’t it?

Faculty_06: That would totally help. [...] 

Faculty_03: [I]t would take a while for a student to come, no matter what
student, to come up and really study the dataset, and really understand
exactly what every data element is. Where it came from and stuff like that. To
think about some possibilities for use in courses. And things like that. But it
would be a great opportunity for a doctoral student.

This last comment introduces a useful point about any of the personnel SHADi might
utilize, which is that longevity could be an issue, in the sense that master’s students
will in general stay at SILS for only two years, whereas most doctoral students will
stay a few years longer. Thus, it might be advisable for the SHADi oversight team to
put some thought into what kind of turnover rate is acceptable for any spot they aim
to fill, with an eye toward offering them to either master’s or doctoral students
depending on the answer.
Activities services. After technology services and faculty services, the last category of services to be considered is activities services. The idea for SHADi activities arises out of the analogy with common book programs. Comments from common book administrators mention activities like public speaker events, essay contests, blogging, discussion groups, art exhibits, dining events, film events, and in one case a faculty member wrote an original play that students performed. At both UNC and NC State, and as is typical of common book programs in general (Laufgraben, 2006), there is an emphasis for activities planning on the beginning of the fall semester. As referred to above, this is closely tied to faculty freedom, and the programs not wanting to interfere with what goes on in the university classroom. It is also about logistics. As one administrator put it:

Com_Bk_Admin_01: We do front-load a number of the events during the beginning of the fall semester. Really because one, it’s on people’s minds since the discussions happen that first week of classes. The author usually comes on campus the first week. So we do tend to front-load more of the programs in general. It’s also a time when students have a little bit more time on their hands. And they maybe have time to be a bit more engaged.

This is a place where a shared datasets initiative, meant for highly motivated undergraduate and graduate students, begins to differ from a common reading program, which is designed for teenagers as a bridge into college life. The SHADi proposal is much more narrowly focused on a smaller audience (i.e., information and library science students) and on a topic in which most audience constituents are predisposed to be interested (i.e., digital data) than is the case with common book programs, which select general-interest readings and serve broad populations of thousands of students across all academic units. Considering this distinction, the
proposed SHADi program aims to be able to integrate itself more gracefully into its academic environment throughout the entire year than most common book programs would be able to do.

It was mainly SILS students who made comments about SHADi events and activities. Unfortunately, nothing like a consensus seems to have formed around what kinds of activities SILS students would want to attend, or when they would like to attend them. Some students commented that they would be interested in going to tutorials to learn different pieces of data technology, while other students said they would not do this. Some students commented that they would like very short events. Others expressed a desire for day-long workshops. Several faculty and students reacted positively to the idea of having lightning-talk sessions or poster sessions, in which students would be able to present work they had done, either in or out of class, related to SHADi datasets. On the other hand, two students related a story about how, as officers of a SILS club, they had arranged exactly this sort of activity, but no one came to it even after several people had expressed interest. Participants were likewise ambivalent about the point in a semester at which they would like to see various kinds of activities and events. Thus, the data collected for this study seem inconclusive about both the type and the timing of events that a program like SHADi ought to plan for — although across the board both students and faculty seemed pleased with the idea of SHADi having some kind of activities program.

In spite of the general lack of consensus, there are two key points that did come across in comments about activities. The first is that scheduling seems to play
a very important role in activities participation. Most students who said that they would like to attend more SILS events said that scheduling was a major factor in preventing them from doing so. They had other classes or they had jobs to attend during scheduled event times, or events would occur on days that the students typically didn’t go to campus. Some creative thinking might be in order, then, for the SHADi oversight team or for an activities subcommittee, to devise a different approach to these problems. It could be worthwhile to reconsider the basic premise of the typical SILS event, which often involves gathering people in a room at the same time. Perhaps SHADi could come up with asynchronous online learning and sharing activities, or could establish low-key “office hours” that don’t necessarily rely on an event-mentality.

The other key point that came across in students’ comments about participating in activities hearkens back to a common element of the student version of “being significant.” Many students discussed how an activity that is affiliated with a recognized group of some sort can be attractive. Some students mentioned that volunteering with organizations outside of SILS can be significant for students by providing them with practical learning experiences, with potential job leads, and with opportunities to help address big questions. Their comments also suggest that there is value in an organization or even a significant dataset simply lending students its imprimatur, as this student exchange demonstrates:

Student_07: I feel like a lot of times the draw for getting involved with stuff is maybe you’ll volunteer or intern with an organization if you think you might end up working there. Or [...] if you work with this dataset, then you can put on your resume that you worked with this dataset [or] this organization, kind
of thing. That would help. You want there to be some payoff for the participants without necessarily being pay or credit.

Student_04: Yeah, to go off that, either that it’s with a bigger organization that might land you a job, or that you’re contributing to a project that then you could put on your resume and show, like, have a link or something that’s like, this is the project that I worked on, this is what I did on it.

What seems to be at play here, to some degree, is an issue of prestige. There are a few comments from students that suggest, if subtly, that the SILS faculty is a group that has the power to bestow prestige on programs like SHADi and even on specific co-curricular activities. Here students discuss how to increase student attendance at SILS events:

Student_08: So something like a symposium, unless you know you can get people committed to it ahead of time, that’s just, it’s not going to look good.

Student_01: I think getting buy in from instructors is helpful. I think students doing it on their own as a one-off is hard. Even if you want it to be more than a one-off it’s hard.

Student_03: Yeah.

These students then discussed, in glowing terms, a gathering or symposium that a particular SILS faculty member organizes annually around a certain course, which often brings in professionals from outside the school. Students seemed to see “the faculty reinforcing this project,” as Student_05 put it, to be an important ingredient of its success. Another student expressed a desire to be able to work with faculty on figuring out what to do with shared datasets in this comment:

Student_07: I think another concern I would have is like, mentorship with this program? Right now if I heard you could do some stuff with this dataset, I would say I don’t know what to do with a dataset, so I’m going to pass on that. But if it were like, all right, here’s an introduction to things you can do with this dataset […] And also kind of like, I don’t know, you have an advisor
for your master’s paper. If there were someone that you could sort of check in with, to be like, I don’t even if know what I did here is, would be logical?

These issues of seeking symbols of affinity from, and seeking close relationships with, faculty and organizations suggest that whomever might be in charge of planning a SHADi activities program ought to strongly consider the role of prestige, derived from both inside and outside the school, in motivating students to get involved. Looked at another way, a longing for prestige could be seen as a need or a desire that students naturally tend to have. Perhaps SHADi could become a vehicle to help fulfill that need at SILS, with SHADi events and activities being a likely method.
Conclusion

The concluding chapter of this paper is composed of three sections. The first discusses some limitations of the research project. The second envisions a first instance of the SHADi project with some recommendations from the researcher. And the third recaps why an initiative like SHADi is something in which a school like SILS should invest its time, funding, and effort.

Limitations of the research project

By focusing data analysis on participants’ comments, this research project emphasizes the subjective experiences of SILS constituents. There are likely some other factors that should be formally considered in the development of a program like SHADi. These include official SILS documentation (e.g., published degree requirements) and semi-official documentation (e.g., individual course syllabi, committee meeting minutes, any available internal PowerPoint slides, etc.), which deserve their own content analysis. Such work is outside the scope of this research project, but it ought to be performed to some level as part of a proper SHADi planning process.

Another potential limitation of this research project concerns the relationship between the motivation for the program and the foreseen outcome, which is mentioned here briefly by way of acknowledgment. The idea that a
technological or systems innovation can or will change the social and academic dynamics at the school could probably be described as a version of technological determinism. Determinism in general has a long and controversial theoretical history, in areas like economics, history, and sociology. It asks, for instance, whether market needs pull in the innovations they require, or whether supply of new innovations pushes demand for them (Flichy, 2003/2007). This is a fundamental philosophical question that won’t be answered here. But it should be acknowledged that the SHAdi enterprise belongs more or less in the market-pull camp, insofar as it first locates a gap in the “market” of SILS academic services and resources, and then attempts to fill that gap. There is no guarantee what will happen to the “market” once that gap is filled. This project’s strategy is to ask for background information from experts and for collaborative input from constituents in order to make the best possible decisions and recommendations, as part of the innovation process. This gets close to the market-pull conceptual ideal, i.e., the idea “that there generally exist[s] a possibility of knowing a priori (before the invention process takes place) the direction in which the market is ‘pulling’ the inventive activity of producers” (Dosi, 1982, p. 149). Of course in the case of this research project, the way to discover the direction of pull in the SILS “market” has been to ask people. It remains an open question how closely the information derived from data collection, and this researcher’s analysis of that data, actually reflects the “pull” of that “market.”

Lastly it ought to be noted that, especially in terms of this project’s contribution to a global scholarly research community, the project is very specific to the particular context of SILS in 2012. Because of this specificity, it is unlikely that
there is much in the data or its analysis that could be broadly generalized to other populations at other times in other locations. This caveat notwithstanding, this project does imagine that a SHADi-like initiative could be undertaken at many ILS programs or iSchools other than SILS. Any other institution would have to conduct local analysis to determine how a shared datasets initiative could integrate with its own populations, its own technologies, and its own curriculum. This research could likely provide a starting point for the conceptualization of another institution's SHADi-like program, but few of this project's analyses or its conclusions would be directly transferrable.

**Recommendations for a first instance of the program**

In this section, this researcher would like to present a brief proposal for how to organize a first instance of the proposed SHADi program. The first thing to consider is the technology infrastructure, because this will influence the inflection of most other facets of the initiative.

**Technology infrastructure.** For a number of reasons, this researcher believes that using iRODS technology is the most attractive answer to the technology infrastructure problem. This technology is a hallmark of SILS, and providing students and instructors with hands-on experience using the technology should give the school's constituents a stake in the cutting edge research that is happening at SILS. Also, compared to any other technology likely be available, the Lifetime Library represents the most sophisticated technology option. There is very little of the interactivity that students and faculty asked for in other technology
infrastructure options like posting datasets on a website or in Sakai. The Lifetime Library, on the other hand, would present constituents at the school with complex file sharing scenarios that could provide students with practical experience manipulating data objects, creating these objects’ metadata, and managing data security — all within the context of a school-wide collaborative effort. It could also potentially provide an object of study for SILS researchers interested in how university students, staff, and faculty collectively manage these issues in the context of an integrated learning system. In this way the SHADI project could provide a unique opportunity for the Lifetime Library team to study their system in situ. This seems to be a potentially shared goal between SHADI and the Lifetime Library team, as this exchange with a faculty member demonstrates:

**INTERVIEWER:** If a resource like [the Lifetime Library] were to be used, say in the SILS classrooms as an augmentation or a replacement for some of the functions that happen in Sakai — presumably you all have thought about this — what are the stepping stones from where we are now to making that actually happen in the classrooms, or on a general basis?

**Faculty_08:** The only thing that has to happen is the instructor has to become comfortable with the technology.

**INTERVIEWER:** Ok. What are the steps between where we are now and the instructor becoming comfortable?

**Faculty_08:** Those steps mean that I need the instructor to start building a personal collection or a class collection in the Lifetime Library. We talked to individual instructors, who either were using BlackBoard, or Sakai, or some other project. And so we’re trying to lure them in by creating common datasets that might be of interest.

**INTERVIEWER:** Well, that’s what — I wonder if [SHADI] could be sort of a —

**Faculty_08:** Yes. This could, yeah.
The idea of all SILS students and faculty having access to a system, not only on a personal basis but also as part of the typical classroom learning experience, that is being designed and tweaked in-house by SILS researchers presents a powerful message to everyone in the school about investing in shared research and learning, and about the school being a community of scientists, practitioners, and students who all share a common identity and common goals. This alone would potentially recommend the use of the Lifetime Library for SHADi but, in addition, the Lifetime Library is the most sophisticated technology available and is the technology that opens up the greatest number of possibilities for practical learning and for new research. Thus, the Lifetime Library is the solution that this researcher would recommend to the SHADi oversight team.

As was mentioned earlier, one objection that might arise toward using the Lifetime Library in this capacity is potential expense. The word “lifetime” is of course built into the technology’s name. It is not, however, a goal of the current SHADi project proposal that all SHADi datasets and related data be made available to all SILS students for the rest of their lives. Instead, this project is focused on the experiences of students while they are enrolled at SILS. There is nothing about the iRODS technology on which the Lifetime Library is built, as far as this researcher can tell, that requires all users to be lifetime users. Assuming that providing all SILS constituents with a Lifetime Library is prohibitively expensive, this researcher would suggest that a non-lifetime version of the Lifetime Library be set up for the purposes of the SHADi program. For this reason, it would likely be wiser and easier to refer to the technology as iRODS instead of as the Lifetime Library.
**Supporting technologies.** While this researcher recommends iRODS as the primary dataset sharing mechanism, it does not mean that no other systems or technologies can be involved. On the contrary, SHADi should avail itself of any tool it can use to serve its mission and its constituents. Social media such as Twitter, Facebook, Pinterest, Google+, YouTube, SlideShare, or podcasts, to name a few, could all easily be imagined to play a role in SHADi, as could a dedicated website, a page or group of pages on the SILS website, or a blog. There are SILS courses (such as a reference sources course or a user education course) that lack obvious opportunities for faculty and students to interact with shared datasets. Some of these ancillary technologies could become ways that people in these courses could get involved with the SHADi program.

**Dataset selection.** There are three major factors that will likely determine a dataset selection approach. How many datasets should we select? Do they need to be related? And how often will we select them?

This researcher advocates selecting as few datasets as seems feasible (without taking it too far), probably between 8 and 12 — a few small ones, a few medium-sized, a couple larger ones. The fewer datasets exist in a SHADi collection, the more “shared” they have the potential of being, as was discussed in previous chapters. Because there seems to be a cultural tendency toward excess (e.g., when traveling most of us probably tend to over-pack our suitcases rather than under-pack them), erring on the side of fewer datasets might help counterbalance symptoms of dataset sprawl.
Identifying a theme can be helpful to unify the dataset collection, and it introduces a synergy to the program in which various datasets invite conversation, invite comparison, invite participants to look beyond a single digital object toward the intellectual, political and social contexts the data could be used to inform. While library-oriented datasets might be an easy first choice, given that SILS is partially a library school, this researcher would suggest that it is difficult to imagine the SILS community getting excited about things like ARL statistics or Library of Congress datasets. Instead, the UNC 2012-2014 “Water in Our World” theme would be much better at prompting students and faculty to dig deeper and to ask significant questions about all kinds of topics, including but not limited to urban planning, international development, desalination, law, parks and recreation, fluoridation, oceanography, health and hygiene, drought, pollution, water sports, and hydraulic fracturing (i.e., “fracking”) as a controversial way to mine natural gas right here in North Carolina. There are also significant possibilities for collaborating with partners across the university and beyond, if SILS were to share a theme with tens of thousands of other students, faculty, staff, and researchers, as well as with government groups and private enterprises, all across the state.

When it comes to the question of how long a group of datasets should compose the SHADi collection, the best answer for now might be, a few years. If the SHADi oversight team adopts, for instance, the UNC water theme, the datasets in the collection will be topical for at least two years, and they could remain important and interesting for much longer. After the first two-year period of running a SHADi program, SILS would be in a far better position than this researcher is now to say
what the dataset lifecycle ought to look like for the initiative. It may be that older datasets can be archived to make way for new data or a new theme, but that the older ones can remain available for constituents who don’t want to lose them. It might make sense to rotate datasets into and out of the collection one at a time. The school’s collective use of SHADi as an initiative that spans the curriculum, and that reaches into independent research projects, will be far more important in the long run than the question of whether or not old datasets will get deleted. The first few years will be important in setting the tone for the initiative. After the program has become established for the first time, and has served SILS for a couple years, it will better be able to grapple with questions of subsequent generations.

Visible successes. Similar to the question of how many datasets should be incorporated into the SHADi program is a question about how many SILS courses, and which ones, ought to be made targets of the initiative. If the stated goal is to integrate the collection’s datasets into the school’s curriculum, where should this be done? Several comments from faculty members advocated the concept that the integration should be broad, that otherwise it could begin to look like some people aren’t invited to use SHADi. However, a number of faculty members recognized that, as Faculty_05 commented, “you could hit all of the students without hitting all of the classes.” This might be done by focusing on certain core courses that all students take, e.g., Organization of Information, INLS 520. This researcher would suggest, though, that in the early stages of the initiative, targeting all students may not be reasonable. A common thread running through the comments of a few faculty members focuses on the need for what Faculty_01 referred to in a comment as
“visible successes,” small victories that can demonstrate the power and utility of the SHADi program. Here a faculty member expands on this idea:

Faculty_05: I suppose you could start with a group of faculty who thought it was a swell idea. And they could use it in their classes. They could produce it, or collect it, or create it, or whatever. They could use it in their classes and then other people could see from their example what they've done. That’s one possibility. [...] I don’t know. I think you’d have to have a really good illustration.

As with the issue of the number of datasets to include in the collection, this researcher recommends that the SHADi oversight team follow the advice of Faculty_01 and Faculty_05 when it comes to franchising into the classroom, which is to start small, and to focus on clear successes. One thing on which many students commented was that, if the SHADi program seemed to be shoehorned into a course in which it did not fit well, everyone involved would carry around, and begin to spread around, a bad opinion of the initiative. All SILS constituents should certainly be courted for the co-curricular aspects (i.e., events and activities) of the program from the beginning, but it would be better to have slow successful growth into the classroom than to create an embarrassing situation where the program’s support systems cannot sustain a too ambitious initial roll-out.

**Why a shared datasets initiative?**

When it comes down to it, why is something like SHADi useful to SILS? Why should anyone care? The best answer this researcher can provide appeals most directly to an optimistic vision of what the ILS field ought to be. SHADi is meant to act as an ice breaker, as a way to stimulate a conversation among SILS constituents
about the growing influence of digital data in the information professions, about what that data can be used for, and about how the school can improve the way it prepares students to solve problems using it. Participation in SHADi will probably not turn anyone into a digital data expert, but it may help ease the school into a future in which leaders of the information professions will need to be able to think creatively about digital datasets and about the kinds of solutions that might be found in them or between them. If the initiative can do this, which is an open question, it will be because it will have found a way to get the school’s constituents to recognize the fact of their shared attention, and to make a small adjustment in it.

The perceived need for such a change in course, even a minute one, derives from this researcher’s perception that SILS could be asking bigger questions — or at least it could be asking its students to ask bigger questions. Students and faculty, in their comments, reported that they want a SILS education to be significant. Where in the curriculum is this need being met? Where is there an emphasis on creativity and entrepreneurship, on using disciplinary knowledge and understanding to tackle big issues? As a leader in the ILS field, SILS ought to be pioneering programs to answer these questions.

This researcher suggests that there is a certain minimum comfort level that one must have with one’s tools and one’s materials before significant creative problem solving is likely to occur with those tools and materials. In the ILS field, digital data of all shapes and sizes are, increasingly, one of the basic materials, and the computer software and hardware used to manipulate, analyze, and store the
data are one of the field's basic tools. The SHADi initiative’s goal is to help familiarize SILS students with these materials and tools so that the school can better focus on asking significant questions whose answers might involve digital data, so that the array of knowledge and skills that information professionals bring to bear (such as a sophisticated understanding of information needs) can be built into the digital tools and data techniques that ought to be coming out of schools like SILS. All of this cannot be achieved by one initiative over a couple years at one school alone, but the proposed SHADi program, which is aimed at SILS, a leader in the ILS field, represents a concrete and localized way to begin increasing our momentum toward this ideal.
Appendix A. Interview Guide

Opening Remarks (Common to All Interview Guides)

Interviewer Introduction: Hello, my name is Chris Weeg. I’m a master’s student in the School of Information and Library Science at the University of North Carolina at Chapel Hill. I’m conducting this study as a research project toward my master’s degree.

Brief explanatory introduction of the proposed SHADi program:

Basic concept: (1) a committee selects a data set; (2) instructors create course content that refers to the data set; (3) students take the courses, a) learning about the data set and b) learning course content through use of the data set; (4) interested students and faculty attend events tied to the data set; (5) interested students present their in-class and/or extracurricular work related to the data set in a voluntary year-end symposium/lab.

Two major goals: (1) to incorporate data management and analysis skills into the SILS curriculum; (2) to provide a means of contact, across many SILS courses/domains, for students to share, compare, and contextualize their learning.

SILS Faculty Interview Guide

My interest in talking to faculty: (1) to gauge interest in a) the goals and b) the methods of the program; (2) to seek input geared toward creating a preliminary program proposal responsive to faculty needs.

Intro

What is your gut reaction to this concept? Positive, negative, in between?

Overall

What, if any, might be positive outcomes of such a program?
What are some likely impediments to initiating this program?
What kind of planning do you imagine would need to happen to initiate a SHADi program at SILS?
How does the project mesh with SILS' mission/direction/future? Library science vs. Information science?

Data Management Skills

How do you think the concept of “Big Data” will influence the information professions SILS graduates will find themselves working in?
What are some attributes of SILS that currently prepare students for data-intensive information professions? Courses that are offered? Projects/programs?

In the next few years, how might SILS augment its current offerings in this area?

What sorts of discussions, if any, have there been within the department in regard to emphasizing data/data management/data analysis/data resources in the curriculum?

What strategy, in regard to data-intensive training, do you favor? What should SILS be aiming for?

If a heightened attention is trained on data, is there another area of the curriculum that would have to go out of focus? What is that likely to be? To what consequence?

**Social/Political**

What are factors about this program that make it likely to gain traction among the SILS faculty? Unlikely to gain traction?

Who would need to get on board for a program like this to catch on at SILS? What are their concerns likely to be, relative to the SHADi program?

**Course Design**

What are the major components of a course's design? (assignments, readings, lesson plan, etc.)

What is the process of course design like? What are the steps? Whom is the design "up to"?

How fluid or static is the design of a course? Across time? Between instructors?

How realistic is it to expect SILS faculty to modify course design by incorporating the SHADi program's data set? What factors make it difficult for faculty to do this? What could be done to make it easier? Would a faculty workshop during summer make sense?

How frequently should a data set be selected? Annually? 3 years?

**Policy re: Collaboration/Sharing**

Deciding how much sharing/collaboration is allowable at SILS?

Potential Honor Policy concerns with a program that encourages SILS students to discuss their learning with each other? How to address these concerns?

Scenario: Student A takes course Z, does assignment Y using the SHADi datasets, presents on findings/work at SHADi symposium. Student B attends symposium, witnesses student A's "approach" to assignment Y, later enrolls in course Z. Is this "cheating"?

**Data Set**

Useful criteria for choosing an appropriate data set?

How might the data set be incorporated into course work?

Can you think of one or two potential nominee data sets?
Courses Not Taken
Perspective of student. Assume 3 types of courses: courses interested in (A); subset of those, courses taken (B); courses not very interested in (C). What do you think, in principal, of trying to create an opportunity for students to engage with the courses designated as A-minus-B? Why?
How might this be achieved?
What do you think of the SHADi program as a way to do this? Other ways?
To what degree would you be willing to guide independent studies focused on the data set?

Closing
Are there any points you would like to make that haven’t been raised yet?
Anything you can offer that I would probably want to know, but that I haven’t brought up?

Common Book Expert Interview Guide
My interest in Common Book programs: (1) to discover to what degree the analogy between Common Book programs (the initial inspiration for my project) and the proposed SHADi program holds; (2) to learn from others' experience about best practices for creating and administering a successful Common "X" program.

Overview
How does the Common Book program work at your institution? Major component? Is the reading required? Technically required? Effectively required?
I'm a student -- what do I do/see? I'm a faculty member -- what do I do/see?
Other groups?

Starting the program
Why start it?
Who was involved? Groups? Individuals?
People's reactions? Any skeptical or fretful response?
Research & planning? How did you know what to do?
Changes since then?

Choosing a book
Who is involved?
Criteria for choice?

Planning & running events
What sorts of events? In-class vs extra-curricular? What events get people involved? What events don't people like?
Problems or issues
What sorts of problems or issues has your institution’s program encountered?
People not reading? Administrative inertia? Cost?
Does anyone oppose the program? Actively? Passively?

Advice for another institution
What sorts of advice would you give regarding:
Should an institution initiate a program? Why? Why not? Signs that I should/shouldn't initiate a program?
What to look out for – likely difficulties, quagmires, tough choices? Whom do I need on my side?
Any thoughts on using an object other than a book?

Closing
Are there any points you would like to make that haven’t been raised yet?
Anything you can offer that I would probably want to know, but that I haven’t brought up?

Data Services Expert Interview Guide
My interest in talking to data services experts: (1) to understand something of the nature of data sets; (2) to discover how data, as a type of information resource, is currently used in the university; (3) to identify criteria for selecting a data set; (4) to ask for advice from experts in the field about program administration and the pedagogical potential of data sets for information and library science programs.

Overall/Intro
Any preliminary advice/comments that come to mind from this introduction?

Overview of data needs/use at your institution
What are the components(objects)? Databases? Structured/unstructured data?
Software? Hardware?
Who are the major constituents? Faculty? Staff? Students? Grad/undergrad?
Outside researchers (from other institution)? Other? Unknown?

Uses of data
Completing assignments?
In the classroom? Do courses point students toward data resources/services? In what ways? Examples? Can you imagine classroom uses not currently utilized? Examples?

What are obstacles to use of data services/resources?
What obstacles do people come to you with?
Software? Compatibility? Cost? Processing power?
Hardware? Special controls/displays?
Finding the "right" data?
Analysis? Do you need to know statistics? Do you need to know computer programming?
End-use? Is it easy/difficult to present data/findings from data?
Access? Cost? Legal/copyright? Time? (computers slow, etc.)? Storage?
Other obstacles? What problems do you encounter/notice?

Data sets
What IS a data set? Are there different types? Is there a taxonomy?
Are some data sets good/bad for beginners? What contributes to this?
Should the Common Data Set program try to "get" the data set, or “point to” it?
What factors should go into choosing one for the SHADi program?
What data set(s) would YOU use to ease students into thinking about data services/resources, letting them explore?
Appendix B. Focus Group Discussion Guides

Opening Remarks (Common to All Focus Group Guides)

Moderator – Introduce yourself. Explain that this focus group is meant to help gauge the feasibility and desirability of initiating a SHADi program, which you’ll describe in more detail later, but in general highlights a single data set for use in a variety of courses, used a) to provide students hands-on interactions with data, and b) to increase communication across a variety of SILS courses.

SILS Faculty Focus Group Guide

Opening

(1) How long have you been teaching college courses, and what is a course you enjoy teaching or would like to teach at SILS?

Introductory

(2) In what ways, if at all, have you used data/data sets in courses that you teach? In-class examples? Assignments? Other ways?

(3) Are you aware of, and can you describe ways other SILS instructors have used data/data sets as resources for a course? Non-SILS courses/instructors?

Key

[Moderator -- describe the premise and major intents of the SHADi program: Select a common object of inquiry (here, a data set) that can be examined or manipulated in a variety of courses, invite students to do coursework "on" the common object, coordinate events (guest speakers, tech tutorials, lab/symposium) that provide vehicles for discussion and shared learning between students across classroom boundaries, give them hands-on experience working with the common object (here, a data set).]

(4) How would you react if you were invited to, on your own initiative, incorporate the selected data set into a course (e.g., in-class lessons, assignments, etc.)? Why?

(5) How would you react if you were invited to use materials developed by the SHADi program administrator(s)? Why?

(6) How would you react if you were invited to co-create materials (for use in a course you teach) with one or more SHADi program administrators? Why?

(7) What unique benefits, if any, do you see accruing to SILS students from the hands-on data interaction aspect of the SHADi program?
(8) How well would the SHADi program work to promote the kind synthesis that weaves different courses into an "education"? Can you imagine other positive outcomes? What are they?

(9) What are some likely points of failure in trying to initiate a program like the SHADi program?

Closing

(10) If a group of students wanted to initiate a program of any type that is meant to reach into and influence curriculum or course design in some way at SILS, what advice would you give them?

(11) What's one important thing that I should leave here knowing, but that hasn't been made clear so far?

SILS Master's Students Focus Group Guide

Opening
[Moderator – give examples of “activities”: Brown Bag lunches, guest speakers, SILS-related employment, clubs, etc.]

(1) Are you on the LS or IS track, and what are some examples of SILS activities that tend to interest you?

Introductory
[Moderator -- briefly describe Common Reads program concept]

(2) What is your impression, in theory or from experience, of the Common Reads or One Book programs that have become popular at colleges and universities? What are some likely positive outcomes? What was/would be frustrating about such a program?

Key
[Moderator -- describe the SHADi program concept in more detail: voluntary, single data set, for use in coursework and in-class lessons, hosted events of various types]

(3) What is your gut reaction to this concept? What are some positive potential outcomes? What is the worst thing about this concept? What are some likely problems that might come up?

[Moderator -- SHADi major intent #1: vehicle for discussion/sharing of learning]

(4) Would SILS benefit from such a vehicle? Why or why not?


[Moderator -- SHADi major intent #2: vehicle for hands-on interactions with a data set]
(6) What kinds of hands-on interaction with a data set would you like to see? Have you had data-oriented experiences in/out of class you’ve found interesting? Examples?

(7) Are there likely obstacles to try to avoid? Contingencies to prepare for? Such as? Software issues?

(8) What could be learned from interacting with a data set?

Closing

(9) Is there a better way to support the major intents? Such as what?

(10) Is there any point you would have wanted to raise but didn't get the chance to? Anything we missed?
Appendix C. Code Definitions Used in Data Analysis

1. **CmBk gen desc.** A general description of what the common book program at your institution does or is.

2. **CmBk events.** What kinds of events are part of the common book program at your institution?

3. **CmBk faculty.** What kinds of involvement does faculty have with the common book program?

4. **CmBk history/origins.** How did the common book program arise at your institution?

5. **CmBk evolution.** How and why has the program evolved over time at your institution?

6. **CmBk selection.** How books are/have been selected.

7. **CmBk into courses.** Ways that the common book readings and events are integrated directly and specifically into courses.

8. **CmBk prog eval.** Regarding the assessment or evaluation activities the common book program engages in.

9. **CmBk difficulties.** Challenges that the common book program has encountered, and the program's responses to those challenges.

10. **CmBk prog resources.** Discussions surrounding the types of resources (e.g., staff, funding, infrastructure) the common book program receives, needs, would like, etc.

11. **What is a dataset?** Comments revolving around the problem of determining what kinds of components would possibly comprise the SHADi collection.

12. **General positive.** Vague or "in general" positive comments about the SHADi concept.

13. **Data selection.** Comments aimed at how datasets might get selected, or by whom, etc.

14. **Data faculty.** Issues surrounding faculty involvement with SHADi.

15. **Data tech.** What kinds of software and hardware are needed, wanted, useful, etc., for making use of datasets?
16. **Data usage.** Discussion of current or historical patterns of data usage at your institution.

17. **Data in courses.** Regarding the real or potential integration into courses of datasets, whether or not the comments are related to SHADi.

18. **Big Data.** Regarding the reputation of data and evidence-based movements in popular consciousness or in the academy as a rising star.

19. **Data literacy.** Regarding students learning to find, select, judge, and interpret data/datasets, and the knowledge they need for this, including statistics knowledge.

20. **Admin via courses.** Comments pertaining to the possibility for SHADi to be shaped or administered collaboratively by students in SILS as a practicum.

21. **Getting going.** Comments about how to ease into a working SHADi program.

22. **Likely impediments.** What are some foreseeable problems?

23. **SHADi resources.** Discussion of the types of resources (e.g., funding, prestige, staff time, infrastructure) the SHADi program may require from SILS, ITS, or some other patron.

24. **Lifetime Lib.** Regarding the Lifetime Library.

25. **Archive SHADi.** Regarding archiving SHADi for historical/documentary reasons, or turning SHADi (and its archive) into an object of study.

26. **Be significant.** Comments pertaining to the idea that the SHADi program can or should be used, not as pedagogical fodder, but as a way to contribute to and engage with big issues and questions.

27. **Social/Academic hub.** Regarding the concept of SHADi being a way to interconnect SILS courses (cf. The Block curriculum), and SILS constituents.

28. **SHADi Events.** Discussion of the possibility for, or the type of, co-curricular events and activities that could be associated with SHADi.

29. **SHADi eval.** Regarding how to assess the SHADi project.

30. **Mesh w/ SILS?** How does the SHADi program mesh or not mesh with the goals, directives, "tracks", and curricula of SILS?

31. **DB curric.** Comments pertaining to the collaborative planning that has gone into courses-sequence curricula such as the databases series, the archives series, the ARM concentration, etc.
Appendix D. Occurrence of Codes/Categories

This paper aims to offer an analysis of the data that were collected from SILS constituents and topic experts, addressing in turn each of SHADi’s four program components with the comments that were made by interview and focus group participants. The descriptive codes that were used to mark up the transcripts constitute the entry point for this analysis.

Table D-1. Data-related descriptive code occurrence by data collection groups. Groups with zero occurrence are omitted.

<table>
<thead>
<tr>
<th>Category</th>
<th>Faculty interviews</th>
<th>Faculty focus group</th>
<th>Student focus groups</th>
<th>Data expert interviews</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin via course</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Be significant</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>Big Data</td>
<td>1 1</td>
<td>2</td>
<td>1 1</td>
<td>1 2</td>
<td>11</td>
</tr>
<tr>
<td>Data usage</td>
<td>2 1</td>
<td>3 2</td>
<td>1 4</td>
<td>2 1</td>
<td>12</td>
</tr>
<tr>
<td>Data tech</td>
<td>3</td>
<td>1 2</td>
<td>1 4</td>
<td>2 1</td>
<td>13</td>
</tr>
<tr>
<td>Data literacy</td>
<td>6 5 3 5 5</td>
<td>2 1</td>
<td>4 2</td>
<td>4 1</td>
<td>39</td>
</tr>
<tr>
<td>Data faculty</td>
<td>6 6 4 7 5 3 5 3 2</td>
<td>4 4 1</td>
<td>3 2 3 1 1</td>
<td>55 4 4 1 1</td>
<td>65</td>
</tr>
<tr>
<td>Data selection</td>
<td>4 9 5 9 3 6 5 4</td>
<td>7 7 1</td>
<td>2 1 2 1 1</td>
<td>27 4 4 1 1</td>
<td>31</td>
</tr>
<tr>
<td>General positive</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Getting going</td>
<td>6 4 2 5 3 2 1</td>
<td>4 2 1</td>
<td>1</td>
<td>31 4 2 1 1</td>
<td>22</td>
</tr>
<tr>
<td>Lifetime Lib</td>
<td>15 2 1 1 1</td>
<td>1 1 1</td>
<td>1</td>
<td>22 1 1 1 1</td>
<td>33</td>
</tr>
<tr>
<td>Likely prob</td>
<td>5 3 1 3 1 2 6</td>
<td>1 5 4 1 1 1</td>
<td>1</td>
<td>32 5 4 1 1 1</td>
<td>6</td>
</tr>
<tr>
<td>Mesh w/ SILS?</td>
<td>4 8 1 4 2 4 1 5</td>
<td>1 3 3</td>
<td>33</td>
<td>3 4 1 1 1 1</td>
<td>6</td>
</tr>
<tr>
<td>DB curric.</td>
<td>1 1 2 2</td>
<td>1</td>
<td>2</td>
<td>1 2 2 2 2</td>
<td>24</td>
</tr>
<tr>
<td>SHADi eval</td>
<td>1 1</td>
<td></td>
<td></td>
<td>1 1</td>
<td>14</td>
</tr>
<tr>
<td>SHADi resources</td>
<td>3 2 2 4 1 3 2</td>
<td>2 4 1</td>
<td>24</td>
<td>2 4 1 1 1</td>
<td>8</td>
</tr>
<tr>
<td>Soc/Acad hub</td>
<td>1 1 2 2 1</td>
<td></td>
<td></td>
<td>9 1 2 2 1</td>
<td>12</td>
</tr>
<tr>
<td>SHADi Events</td>
<td>1 1 1</td>
<td>6 4 1</td>
<td>14</td>
<td>1 6 4 1 1</td>
<td>14 37</td>
</tr>
<tr>
<td>Archive SHADi</td>
<td>1 1 4 1 1</td>
<td></td>
<td></td>
<td>8 1 1 1 1</td>
<td>16 24</td>
</tr>
<tr>
<td>What is a dataset</td>
<td>2 2 1 3</td>
<td></td>
<td></td>
<td>16 2 2 4 1 1</td>
<td>14 24</td>
</tr>
<tr>
<td>Totals</td>
<td>41 45 49 51 32 34 37 34 47 37 24 14 23</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table D-2. Common book program-related descriptive code occurrence by data collection groups. Groups with zero occurrence are omitted.

<table>
<thead>
<tr>
<th>Code Type</th>
<th>Faculty interviews</th>
<th>Student focus groups</th>
<th>Common book administrator interviews</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>CmBk difficulties</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>CmBk events</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>CmBk evolution</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>CmBk faculty</td>
<td>1</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>CmBk gen desc</td>
<td></td>
<td>2</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>CmBk history/origins</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CmBk into courses</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>CmBk prog eval</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>CmBk prog resources</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>CmBk selection</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Tables D-1 through D-3 provide an initial summary view of the way these codes were applied to the interview and focus group transcripts. Tables D-1 and D-2 depict the occurrence of the codes by data collection category (e.g., faculty focus group, common book administrator interviews, etc.), with Table D-1 showing data-oriented codes, and Table D-2 showing common book program-oriented codes. The two code types have been separated here into two tables merely for space considerations; in each table, any data collection category was omitted if it had zero occurrence for all rows in that table.

Table D-3 depicts the occurrence of code co-occurrence in the study’s transcripts, that is, the overlap of multiple descriptive codes on any particular portion of text. The interpretation that is being used here is that codes which co-occur are likely to be related in some way. The table represents most values both as numerals and also as shades of gray, in order to aid visual analysis, and in keeping with the qualitative interpretation of these numbers. Values below 5 have no shading (zero is shown as a blank square); values from 5 to 9 are shaded light gray; values from 10 to 14 are shaded dark gray; and values of 15 and greater are shaded black. By triangulating between these records of code co-occurrence and code usage attributed to populations, it is possible to find hints of meaning, useful as leads that can be followed up by closer analysis of the transcripts.

An example of one of these hints might be related to the attribution of the “data usage” code. This code was used to mark transcript excerpts dealing with the ways students and researchers make use of an information organization’s data resources and services. Not surprisingly, we see in Table D-1 that data experts were the only group to whose comments the “data usage” code was attributed. And in Table D-3 we see that this code co-occurs most highly with comments about data technology and data literacy. This
Table D-3. Descriptive code co-occurrence. Codes related to common book programs were omitted due to lack of space. Values less than 5 are shown with no shading (zero is blank); values ranging from 5 to 9 are shaded light gray; values ranging from 10 to 14 are shaded dark gray; values 15 and greater are shaded black.

<table>
<thead>
<tr>
<th>Admin via course</th>
<th>Be significant</th>
<th>Big Data</th>
<th>Data usage</th>
<th>Data tech</th>
<th>Data literacy</th>
<th>Data faculty</th>
<th>Data selection</th>
<th>Data in courses</th>
<th>General positive</th>
<th>Getting going</th>
<th>Lifetime Lib</th>
<th>Likely prob</th>
<th>Mesh w/ SILS?</th>
<th>DB curric.</th>
<th>SHADi eval</th>
<th>SHADi resources</th>
<th>Soc/Acad hub</th>
<th>SHADi Events</th>
<th>Archive SHADi</th>
<th>What is a dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adm. via course</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Be significant</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Big Data</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data usage</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data tech</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data literacy</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data faculty</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>13</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data selection</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>13</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Data in courses</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>26</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>General positive</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Getting going</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>19</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Lifetime Lib</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>9</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Likely prob</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>13</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mesh w/ SILS?</td>
<td>1</td>
<td>6</td>
<td>13</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>13</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DB curric.</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHADi eval</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHADi resources</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Soc/Acad hub</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHADi Events</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Archive SHADi</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>What is a dataset</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
SHADi program should include the consideration that the program’s success might depend on how well SHADi prepares itself to grapple with issues of technology and data literacy, perhaps by treating them as likely points of failure: instructors and students who cannot open and manipulate digital data files, and who cannot fathom what SHADi’s datasets mean, are unlikely to find much value in sharing these objects of frustration.
References


School of Information and Library Science. (2010). Bylaws, policies, and procedures of the School of Information and Library Science, the University of North Carolina at Chapel Hill. [Unpublished document].
