From Writing and Analysis to the Repository: Taking the Scholars' Perspective on Scholarly Archiving

Catherine C. Marshall
Microsoft Research Silicon Valley
1288 Pear Avenue
Mountain View, CA 94043
+1 650 693 1308
cathymar@microsoft.com

ABSTRACT
This paper reports the results of a qualitative field study of the scholarly writing, collaboration, information management, and long-term archiving practices of researchers in five related subdisciplines. The study focuses on the kinds of artifacts the researchers create in the process of writing a paper, how they exchange and store materials over the short term, how they handle references and bibliographic resources, and the strategies they use to guarantee the long term safety of their scholarly materials. The findings reveal: (1) the adoption of a new CIM infrastructure relies crucially on whether it compares favorably to email along six critical dimensions; (2) personal scholarly archives should be maintained as a side-effect of collaboration and the role of ancillary material such as datasets remains to be worked out; and (3) it is vital to consider agency when we talk about depositing new types of scholarly materials into disciplinary repositories.

Categories and Subject Descriptors
H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *system issues*, *user issues*

General Terms
Design, Documentation, Human Factors

Keywords
digital archiving, collaboration, scholarly publishing, qualitative study, personal information management, scholarly repositories

1. INTRODUCTION
Considerable attention has been given to the construction, maintenance, sustainability, and interoperability of scholarly archives [11, 16]. DSpace, Fedora, Eprints, arXiv, and other institutional and disciplinary repositories and repository infrastructures are aimed at an audience of research libraries that serve scholars and their academic institutions [7]. Less attention, however, has been paid to how the publications are created and assembled prior to their deposit into one of these repositories and how scholars maintain their own archives. This paper explores a focused subset of upstream practices associated with collaborative authoring, reference use, and the informal creation of personal archives. We assume the perspective of a specific community of individual researchers and small groups of collaborators who write papers together.

JCDL '08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

Why look at the publications before they are ready for our more formal institutional archiving efforts? After all, research and writing are highly variable creative practices and it seems that there is little connection between the conventional set of tools—the text editors, analysis applications, and collaboration infrastructures—and the repositories that store the finished products. By the time publications are deposited, they are in a small number of standard formats (e.g. PDF) and any associated datasets are similarly standardized and documented via metadata.

Yet, a number of archiving projects have noted problems at ingest time—everything from viruses in the deposited files [1] to unwillingness to deposit anything without additional impetus (see, for example [6]) to difficulties in documenting scientific datasets [3, 4]—and there are intimations from these findings that it might be wise to look further upstream. Furthermore, from the point of view of the researchers and scientists themselves, institutional archiving arrives on the scene late in the process; the deposit of publications and datasets is an afterthought to the actual work, the research and writing. What would make archiving more integral to the entire process? What does scholarly archiving look like today from the scholar's perspective? How can normal collaborative interactions be used to improve repository quality?

There is a considerable body of work on writing as a cognitive process (see for example [9]). Besides trying to characterize the cognition involved in writing, some of this research has been aimed at authoring tools that expose the structure of the written product [2]. This study is not so much concerned with the cognitive aspects of writing; rather, it is focused what might be considered the social, technical, and mechanical aspects of scientific scholarship: How files and datasets are exchanged, replicated, and moved among computers and collaborators while the work is in progress; how and why versions are maintained; how a set of collaborators comes up with related work; and how researchers keep their own archives of publications and ancillary material complete and up to date.

The investigation described in this paper was performed in service of the design of a scholarly writing application that uses peer-to-peer file sharing. The application supports the filtered replication

of collections so researchers may share files with themselves across multiple devices (for example, replicating files between a desktop and a laptop) and so they may share files directly with their colleagues, bypassing intervening servers.

In this paper, I take the long view of the field data and discuss findings in four areas: (1) collaboration and writing; (2) file storage and management; (3) maintaining and extending bibliographic resources; and (4) personal archiving of scholarly material. After presenting the study's findings, I discuss their implications for collaborative information management, personal scholarly archives, and institutional and disciplinary repositories.

Because this study examines the practices of Computer Scientists, we might wonder if the researchers represent a best case scenario for the use of scholarly archiving systems; the study's participants understand the underlying technology and are acutely aware of the dangers and challenges associated with the long-term storage and retrieval of digital data. Yet the problems they report in this study are surprisingly recognizable and universal [13].

2. STUDY DESCRIPTION

This paper reports the results of a qualitative field study of the research, writing, and archiving practices of fourteen Computer Science researchers working in a corporate research setting. The researchers work in five overlapping (but distinct) subdisciplines: Algorithms and Theory; Distributed Systems; Security and Privacy; Software Tools; and Web Search and Data Mining. The fourteen participants identified themselves as belonging to either one or two subdisciplines (Table 1); participants were selected to reflect the makeup of the lab at the time of the study.

Table 1. Breakdown of the study participants' subdisciplines

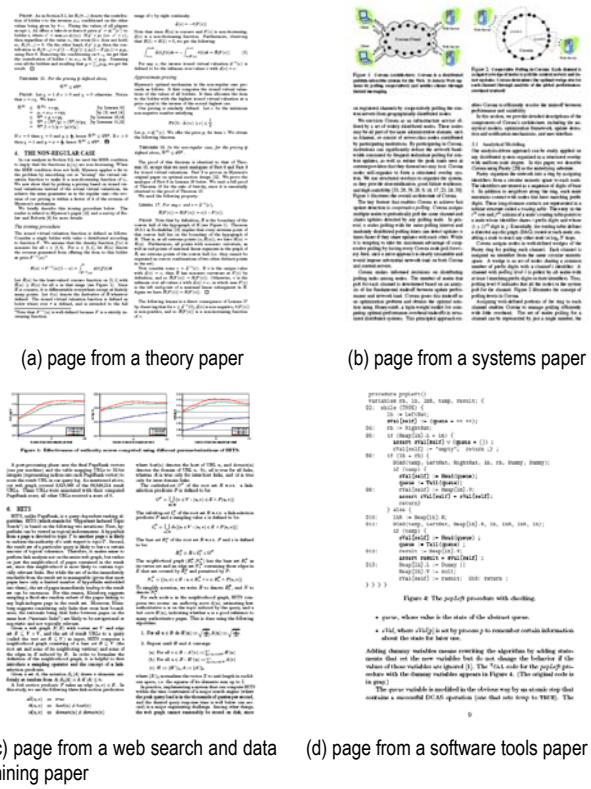
Subdiscipline	*# of participants*
Algorithms and Theory	3
Distributed Systems	10
Security and Privacy	5
Software Tools	3
Web Search and Data Mining	3

I conducted fifteen semi-structured, open-ended interviews (one researcher was interviewed twice to take advantage of a newly-completed collaborative authoring experience) and observed on-going collaborations over the course of six months. The interviews ran from 45 minutes to over an hour and a half, depending in part on the type of research the participant did; the interviews tended to be longer if the participant worked with large, complex datasets (as the Web Search and Data Mining researchers did) or developed and evaluated substantial pieces of software (as the Distributed Systems researchers did). All interviews were recorded (audio) and supplemented with digital photographs (for screen capture). The interviews were transcribed and the transcripts were carefully analyzed.

To ground the interviews in specific artifacts and collaborations, each interview centered around one or two recent papers—usually the last paper the researcher had written, unless it was felt to be not representative of the researcher's oeuvre—and the set of co-authors who wrote it. The interviews expanded from discussion of the recent paper into bibliographic practices, how references were gathered for this paper and for other recent papers. We then discussed one or two of the researcher's older papers. Because it is difficult to recover the minutiae of a long-ago authoring process, for the older material, we focused on what was kept, where it was kept, and how it was kept. We also discussed data loss at this point and concluded by talking about the researcher's general archiving practices. In half of the cases, I was able to interview multiple authors of the same publication.

Publication genres differ: Figure 1 shows a distinctive page from papers representing four of the five subdisciplines (Security and Privacy papers had similar visual elements to Algorithms and Theory papers). These structural differences (the presence of certain kinds of figures, e.g.) were used as cues in the interviews.

Participants were in different phases of their careers (as indicated by their job titles, which included Research Developer, Researcher, Senior Researcher, and Principal Researcher). Seven were early in their careers, just a few years out from their PhDs. The other seven were mid- to late-career and had changed institutions multiple times. This diversity of experience enabled us to check for changes in research, publishing, and archiving strategies, and it ensured that participants had a significant legacy of research and publication, some of which had been completed under other institutional or corporate auspices. Regardless of where they were in their careers, all participants had external collaborators, usually at academic institutions.



(a) page from a theory paper

(b) page from a systems paper

(c) page from a web search and data mining paper

(d) page from a software tools paper

Figure 1. Visual characteristics of publication elements

All participants published regularly; most regarded publication as an important part of what they do (although some focused more on developing intellectual property or research code that could be

transferred to product teams). Even the more junior researchers had significant publication records; they averaged close to 25 papers apiece. Several of the most senior people interviewed had over 100 peer-reviewed publications to their credit. Thus all participants were experienced paper-writers and researchers.

In the process of interviewing the participants, I collected relevant artifacts such as Web-based publication lists, *curricula vitae*, and the publications themselves; I also examined datasets, slide decks, and documentation for the commercial software they used (e.g., MATLAB, Beyond Compare, and Source Depot).

3. FINDINGS

The findings are divided by the four areas of investigation: collaboration and writing; file storage and management; maintaining and extending bibliographic resources; and personal archiving. Because there has been significant qualitative research in the CSCW community aimed at describing some of these activities (e.g. [5]), I narrowed my sights on what is necessary to answer the research questions. Although I begin with a broad description, I make an effort to focus closely on the practices and artifacts relevant to maintaining personal archives and contributing to institutional repositories.

3.1 Collaboration and Writing

I begin by setting the stage, describing how researchers write papers together and discussing how the participants use different devices and the organization's technological infrastructure. By understanding the activities central to writing, we can start to see how the upstream research practices feed into the larger scholarly sphere. We can also begin to get a sense of how the materials are shared—where authoritative copies are, both within a particular researcher's set of devices and how they move among a set of collaborators. Finally I discuss the technology environment.

Roles in writing. Naturally, every paper represents a shifting set of roles and responsibilities for authors. Each author brings something different to the table—for example, knowledge of related work in an area, analysis techniques and methodological commitments, and even skill at using various applications—and each author has different preferences and writes under different circumstances. For example, some study participants prefer to coordinate the writing by parceling out sections; others prefer to write an entire first draft by themselves, and not cede the text to their co-authors until it is fully formed.

Coupled with these preferences are the material circumstances of the collaboration, how and where the collaborations take place. Participants who are in the same place as their co-authors often said they prefer to write shoulder-to-shoulder, sitting in front of a display and working together closely to formulate the text. If they are not co-located, synchronous collaboration might take place over the phone, with drafts passed via email. Asynchronous collaboration meant draft-passing or per-section assignments. The research was often split into skill-based pieces: system builders worked with evaluators, for example, and this division was reflected by the writing assignments. Naturally modes of writing are not mutually exclusive: a paper may start out as written in parts, evolve into a draft-passing collaboration, and finally transition into a synchronous activity where co-authors sit together to iron out the kinks in a final draft.

How do these characteristics of collaborative authoring bear on our research questions? First, it is important to examine the effects of the different roles that co-authors assume. Even if their skills overlap to a considerable extent, some co-authors are less involved with the writing process than others, and hence may wind up with incomplete sources for the paper for their own personal archives. For example, consider this story one participant told about working with a group of students from his former institution, which is located on a different continent:

> "The students are writing all of the code and generating all of the results… In the 5 or 6 weeks leading up to the deadline, they will move to spending 100% of their time writing code and we will have regular phone meetings and I will start writing the text of the paper. And then once there's sort of a draft and the results are starting to be clear, then we shift into a second mode. And it ends up with them having final control over everything as the deadline approaches."

At the end of the process, the co-author who was interviewed will not have the final draft unless he asks for it (which he did). He also will not have the full datasets, nor is it practical for him to be able to get them. He did have in hand the subset of the data used to create some of the figures: "Examples and graphs. And overall statistics. ... But not the entire database."

Given the distribution of knowledge and skills, sometimes even the first author has incomplete knowledge of the methodological or analytic details underlying portions of the paper. One participant talked about how he and a co-author had worked on different parts of the paper and even used orthogonal tools: "I have no idea how to use GnuPlot. And I'm sure Norman doesn't have any idea how to use Visio. [laughs] He probably doesn't even have it on his computer." This factor not only dictates who has copies of what (again, the speaker does not have the datasets); it will also become relevant downstream if the datasets are to be deposited in an institutional repository or an author wants to re-use the figures in a presentation.

Communication through alternate channels (including face-to-face meetings and phone calls) may make it harder for a system to record context or document provenance. The work-arounds needed to cope with non-negotiable differences in platforms, infrastructure, or security mechanisms may also result in an incomplete record or incomplete set of files for the archive.

Finally, the interviews reveal opportunities for recording context when writing drives research, as the need to fill in details drives additional experiments. One participant explained:

> "[In] this paper…the data points actually were computed as we were writing the paper, so we just had placeholder figures and filled in the points."

This dialectic between text and data may provide a valuable window onto the meaning of datasets.

Places and devices. Despite the quest for ubiquitous computing and seamless replication across devices, place has an undeniable effect on what people do—some venues are just more appropriate for writing and research than others. Noise (or lack of it), display size, firewalls, the presence of colleagues—all of these factors play into where participants write. Similarly, different computers may be tied to different activities; they may be dedicated number crunchers, email machines, or backup hosts. These functions dictate where materials are stored.

The places researchers work aren't simply office and home. Some researchers have offices that correspond to dual affiliations (academic and corporate); these offices duplicate some resources (e.g. physical storage, servers, or email), and not others (e.g. colleagues). Three participants worked during long train commutes, which required them to be relatively self-contained on laptops. Finally, conferences and other travel are legion in most researchers' lives—but are they venues for writing, or are they venues for human interaction? A participant described his travel this way: "[When I travel] I tend to my email. I don't particularly write. … I view the main purpose of a conference is really to be an opportunity to socialize with people in the field."

Researchers may not work on the same things everywhere; for example, a second office may be used for meeting with students. Some venues are only appropriate for lightweight activities—reading and marking up a draft or collecting new references to read, but not writing or analyzing data. Furthermore, it may not even be *possible* to do everything everywhere because resources (e.g., large datasets) are missing.

Thus, at the end of any writing process, it is possible that no single collection of files localized to a PC or server represents the 'ground truth' for the publication—any given set of files may be either incomplete (lacking a dataset, for example) or not up-to-date (the penultimate draft).

With discussions of place come pragmatic issues of security and the protection of intellectual property. Firewalls can have a profound effect on how work is done. Researchers who normally exchange files through a repository or through the file server shift to email when they are outside of the firewall: "I don't like going through the firewall… it doesn't always work…I prefer email, when I'm traveling, to the file system." Documents that reveal intellectual property made participants wary of copying full collections outside the firewall, even if they worked in more than one office:

> "Some of the files reside on servers at [my company], some on servers at [my university]. For my old papers, the ones I did before being a regular employee, they are typically in [my university]. Though sometimes I have moved over a copy here to have access to them. Typically I know there are legal issues with that."

Tools. Despite an apparent homogeneity within the local organization ("We live in a LaTeX world," said one participant), there is considerable discrepancy in the supporting applications that people use. These applications fall into four categories:

- Editors, figure generating tools, and document preparation software: content preparation applications;
- Analysis software used to process data;
- Infrastructure software such as email and Source Depot used to support the collaboration itself; and
- Custom software, often in the form of scripts and other small programs, used specifically to produce specific results for the paper.

The heterogeneity of content preparation tools—including emacs, LaTeX, PowerPoint, XFig, Visio, JGraph, Word, and other packages too numerous to mention—often makes the generation of publications from their constituent parts difficult, especially for any single co-author. Thus, heterogeneity reinforces roles in the collaboration; the person who dictates tool use will continue to be the one who generates components for the paper using that tool.

3.2 Storing and Managing Materials

Participants expressed a full spectrum of views of what should be stored with the body of a publication: papers are necessarily more than the constituents of the actual publication. For example, they may rely on supporting material such as datasets, logs, code, and so on; items such as presentations may be based on the publication (and may even share objects like graphs). Thus at one end of the storage spectrum is the perspective that the collection should include the items necessary to assemble the paper and support its production, possibly including data files used to generate the results, *and nothing more*. At the other end of the spectrum is the perspective that the collection should include anything loosely related to the paper (but possibly tangential to its production) such as presentations on the topic and datasets that represent unpublished related investigations. In practice, none of the informants in this study inhabit the extreme ends of the spectrum, even if they characterized themselves as doing so.

What the perspectives have in common is that it is important for the set of files to be *complete* (that is, all the pieces are there to assemble the paper) and *up-to-date* at the end of the writing process (that is, final versions of each part have been obtained). To some researchers, it is also important that this collection has been tidied up. Finally, it is important to know which version of each element is authoritative, because multiple ambiguously named versions are often stored together.

Thus there are two aspects of collection management that I will discuss because they have a significant effect on sharing and on personal archiving: versioning and working with data.

Versioning. Research papers may undergo many revisions, some minor, some major, during the publication cycle; when there are multiple authors involved, revision may be simultaneous. Some revisions are uncontroversial: errors are corrected; essential material is added. Others are more contentious: explanations change; new modes of proof are offered; and text is deleted. Each set of revisions has the potential to result in a new version of the publication. The question is, are these versions important artifacts that should be kept after the paper is complete? If so, which ones?

Participants use sophisticated versioning systems for the tools they provide, such as the ability to compare successive versions or to merge two conflicting versions. Secondarily, participants use these mechanisms as a way of creating personal backups to stave off catastrophic losses or simple unintentional deletions. If every save operation results in a version, it is potentially possible to recover a clever turn of phrase that disappeared midway through the writing process. Thus, in its most literal and quotidian use, the main functions of a versioning system are coordinating changes in collaborative work and backing up the content as it changes. Neither use suggests that versions are important in the long run.

Yet there are two other more complicated reasons for using versioning tools. First, participants use them to create *meaningful checkpoints* in the research lifecycle; this provides collaborators, reviewers, and editors a stable document to refer to. It is not unusual for a paper to move through multiple such states of completion: a completed technical report may be shortened into a conference submission, which is in turn revised according to reviewers' wishes and possibly reformatted to meet publication

requirements. This version may in turn be revised and extended into a journal submission, which goes through another cycle of review and reformatting. In this way, it is easy to accumulate six or seven meaningful versions of what may be conceptually a single publication.

Second, participants feel that versions record the development of ideas, a trail that may prove important. But how important? Much of the history and provenance of an idea can be reconstructed from communications media like email, especially when it is combined with intrinsic metadata such as file dates. Thus benign neglect coupled with imaginative interpretation will get you pretty far in reconstructing a publication's history. We are left with an ambiguous assessment of the value of versions.

Managing data. For 9 out of 14 participants, creating, gathering, analyzing, and presenting data is a fundamental part of their research and publications. Datasets present the most problematic part of storing, managing, and archiving a publication, especially when the datasets are large, changing, or difficult to recreate.

First, it is important to understand how datasets are related to a publication; that relationship may give us some insight into how the data should be stored. Certainly many publications in the systems and machine learning/web search subdisciplines are data-driven—results of experiments or simulations reveal what should be highlighted in the paper. Datasets may be very large (hundreds of gigabytes or even terabytes), labor intensive (hand-tagged training sets), or expensive. In these cases, a given dataset may connect a whole family of related publications. In other cases, the dataset is an important component of a single publication; it may represent the simulation used to evaluate a system, for example. Finally, sometimes drawing graphs and making charts is illustrative rather than being a revelatory part of analysis. In the most extreme cases, the data does not exist independently; rather, it is just part of the publication source:

> "Occasionally I actually do have a little performance data for this paper and I believe that what I did is I just put the performance data inside the paper source file as comments. Just so they'd always be there if I needed them."

At what stage should the datasets be set aside as archival? There are often huge primary datasets that are the raw material of many different analyses. These datasets may be gathered through a web crawl; they may be created by a simulation of hardware or software functions; they may be system logs that reflect actual use; or they may be corpora of video, text, or some other kind material used to test an algorithm or technique. The actual dataset used in a particular publication may be derived or winnowed down from one of these larger data collections. Is it sufficient to keep the code used to create the derivative form? "They can be regenerated *as long as you maintain the same state of the tools*," one participant said (italics mine).

The trajectory of a given dataset may be complex. Not only may they be derived from much larger datasets, but also, as in other sciences, datasets may need to be massaged, corrected, gap-filled, or otherwise post-processed [17]. A participant described fixing input data for a simulator:

> "This is a case where you can reasonably assume that if you were to take this and make it a real system rather than a simulation, the requests would be aligned. So we just take the requests and round them up to be zero by eight."

At some point downstream, a dataset may be fed through an application to arrive at a consumable form, the graph or plot that goes into the publication; data may be graphed using Excel or GnuPlot, or they may analyzed in MATLAB.

Finally, even if there is consensus about which datasets should be saved and at what points in the analytic arc, there are still questions about *how* to store them so they are in an accessible form (such as comma separated value format) and documented (for no data are self-explanatory). Descriptive metadata—column headings and comments—may get in the way and prevent the dataset from being used. When this happens, participants describe developing scripts to *remove* the explanatory metadata so the dataset can be ingested by tools like MATLAB. In addition to documenting the datasets, it is often necessary to document processing elements such as input parameters or compiler options; the participants who followed complex processing trajectories often kept logs and transcripts, some in email, and some in ad hoc lab notebooks. One researcher even developed a method for maintaining a lab notebook using Microsoft OneNote:

> "For a while I was actually using OneNote as a kind of an engineering log book, where for every experiment I would cut and paste everything into it and things like that. And that actually worked fairly well."

If a dataset is large and is the source for many publications, it may have something of a life of its own—its own directory, its own hard drive, or possibly its own processing hardware. *Large datasets may not be backed up, let alone archived.*

Any discussion about storing datasets raises complementary issues about storing any specialized code used to generate, analyze, clean, or otherwise process the data. This code may be part of the research code, part of the database (as stored procedures), or may exist independently (e.g. as scripts). Depositing research code is ultimately an important part of the entire archiving process, but there is little guarantee that the software can be kept running over time; it is possible to argue that in some cases, publication extracts the value from the datasets and the code. A participant describing how he stores data concluded by saying:

> "Sometimes the programs have to start a complex simulation which runs on many machines doing something. Sometimes I keep those programs around too. So I'm able to run it again. But this is a few years later and the programs don't exist anymore."

Capabilities for storing, documenting and versioning data will ultimately rely on anticipated future use: Will the data be reused as the basis for other, possibly unrelated, projects? Or is the data available so the reader can verify the authors' results? Can the data be used independently from the associated code? Future intelligibility and reuse crucially depends on these capabilities.

3.3 Bibliographic Resources

Keeping up with the field and maintaining a good sense of related work is inarguably a vital part of performing research. Bibliographic practices are examined from the vantage point of maintaining and extending actual resources (as BibTeX files). Since there is extensive work in this area, the discussion is confined to what is necessary to answer the research questions and characterize this group of researchers.

Maintaining local bibliographic resources. The ability to maintain local bibliographic resources is important. Participants often cited LaTeX's bibliographic capability—BibTeX—as reason enough to use it to prepare publications. LaTeX's bib files are a cumulative investment: participants build them up over time and use them as a type of intellectual bookkeeping to keep track of what they have read, why they have read it, and where they found it (as a URL). In fact, this bookkeeping function tended to prevent participants from merging their co-authors' entries into their main personal bibliographic resource:

> "It often happens that when I write a paper, I contribute some of the bibliographic entries, my coauthors contribute some. Then they start from my bib file, but then it gets modified. I never bother merging those back into my main bib file."

Because sources are gathered from a mix of authoritative and non-authoritative web sites, participants find it important to keep URLS for references. Some participants also store the PDFs with the other material related to the paper: "Because that way, if I'm working on the paper, I can copy the whole directory to my laptop and I can go back and read the reference papers."

Several participants extend citations with comments, summaries, notes, tags (classifiers), or abstracts. These extensions help researchers remember what was in a particular paper, why they thought it was important, and whether it was good. Abstracts and tags make it easy to find citations again. Some of the extensions that participants describe—especially candid comments about the quality of the work—are regarded as private:

> "Comments are very sensitive you know. Sometimes you say, 'I didn't like this paper.' … Once you give them to somebody, you don't know where they'll end up. …[So] when I give the file to somebody, I have a script which strips out all of the comments."

Bibliographic accuracy varies. Of course, verifying that the reference is the proper one to cite is uniformly regarded as an essential part of scholarship, but participants disagreed on the importance of citation details. Some participants said that they verified aspects of the citations such as page numbers that are dictated by the publication's hardcopy form. For others, bibliographic accuracy is a pragmatic matter: will a reader be able to find the reference using the citation? If so, the formal details like page numbers don't matter. Other pragmatic factors enter the picture too: to reduce paper length, a bibliographic citation may be shortened by using abbreviations and omitting the redundant information from the complete citation. Thus entries in bib files may not only be duplicated, but also may diverge in content.

Participants describe other implicit distinctions among references. References necessary to support a claim in a paper's argument (*peripheral references*) are distinguished from references that are closely aligned with one's own work (*central references*) and references that are foundational in the field (*foundational references*). Currently, this distinction might be realized by whether the paper is kept (as a hard or soft copy). A researcher may retain a copy of a foundational reference to give to others and a central reference because it is important to accurately represent the paper's claims. On the other hand, a peripheral reference may be read quickly on the screen, just to pick out the necessary support for an argument. For example, one participant said:

> "When there is a PDF file I'm citing, for things that are relevant to the heart of the paper, I will make a hardcopy and put it into one of those folders... But for things which are more tangential, I'll just say, 'look at x'."

There is some tension about how BibTeX files are organized. Are they a centralized resource, or are they stored with the paper and implicitly partitioned by topic area? Participants cited some obvious benefits of centralization—every citation is easily kept to-hand and there is only one authoritative entry for each citation. A participant, whose BibTeX file was highly regarded by other participants said:

> "Everything I read goes in this file... So when I find a paper, I write something here and I type some comments, including technical stuff… So this is a pretty big file. It has 40,000 lines. It's definitely more than 1500 papers now."

On the other hand, if BibTeX files are tied to a paper, it is possible to have an up-to-date related work template in a research area. Participants who maintain their BibTeX files this way describe starting out on a new paper by copying over the bib files from the last similar effort. One researcher said, "As papers get written on the same subject, then [the paper] inherits the previous bib file. [I] Just copy [the bib file] from project to project." Participants noted that decentralized bib files contain redundant entries, making it hard to propagate corrections and annotations.

Non-traditional resources. Participants consult a variety of non-traditional references such as blogs and Wikipedia entries. This complicates matters considerably since there are fewer assurances of the stability and authority of such of references.

How do researchers get into the literature when they are writing a paper? One important strategy is for members of the organization to rely on each other and each others' citation lists to make initial forays into new material. One participant expressed it this way:

> "I do a lot of interdisciplinary work and when I'm doing something in an area where I'm not comfortable with the literature, I will ask someone in that field."

New co-authors bring their bib files to the table with them when they enter a collaboration; in some sense, it's their intellectual dowry, a way in which they extend the group's reach. Because of this, sometimes a new collaborator is charged with writing the related work section of a paper.

Given a suitable starting place, link following (e.g. from citations or reverse citations) is regarded as a reliable way for locating new material, especially in an unfamiliar area. "Mostly I find references through the reference lists of other papers," one participant said. Another echoed that sentiment:

> "Once you've gotten into the literature then there are citations. That seems to work fairly well…. You find one or two papers using Google and then go hunting through their citations. Or even just searching on the name of that paper. You get things that cite it. You do forward citation lookups."

Participants considered CiteSeer as a useful resource for doing citation following, but several commented that it less useful now, partly because it is not kept current. They simulate CiteSeer's reverse citation linking by performing their own reverse searches (that is, by putting a distinguishing portion of a paper's title in a search engine and seeing who has cited it). They also use DBLP for traversing among co-authorship relations.

Search plays its most important role by helping researchers find a reference for which they have partial information, e.g. the name or affiliation of one author or all or part of the paper's title. This partial information is usually sufficient to find the reference and to construct the exact citation:

> "If it's a paper that I've cited recently in some other paper, my first reaction might be to go to my bib file for that paper. Otherwise, I might try to find the web page of the likely authors of the paper."

Searches of this sort usually are not performed using specialized search engines (like Google Scholar or Live Academic) or inside particular digital libraries (like ACM DL, IEEE DL, or specific journals or conference proceedings). Rather those resources come into play later for establishing the authority of a downloaded reference or for verifying the credibility of the citation metadata:

> "I will either be aware of the paper and I will use Google to get the exact citation and then go to the ACM library to get the paper. I don't use the ACM library for discovery."

Thus the search becomes a several step process: first the reference is pinpointed, and then it is retrieved from a reliable source such as a publisher's digital library so an authoritative version of the paper can be obtained along with an accurate citation. Participants added that personal web pages are good sources for certain kinds of publications (e.g. those available by subscription only), although they are not regarded as authoritative. In between are quasi-maintained databases like DBLP and CiteSeer.

Coverage, *scope*, *authority*, and *timeliness* are intrinsically linked with the utility of particular online resources. For example, most interviewees choose their search engine on the basis of coverage. Although their in-theory utility is acknowledged, scholarly search services (e.g. Google Scholar and Academic Live) are rejected as research search engines by a substantial portion of the participants because neither their coverage nor timeliness is as good as their non-academic rivals: "I'm not using any of the academic engines. ... Because I'm not trying to find the answer right there; I'm trying to find the page number that contains the answer."

New genres of publications are becoming increasingly important to participants. For example, blogs are cited as a good window into what expert practitioners are doing. This material is not duplicated in traditional sources, yet it is important to consult:

> "This guy has a fantastic blog. He's actually a software architect at Microsoft… and he writes about a lot of issues in data centers... There's a lot of links and powerpoint presentations and stuff. And he blogs almost every day."

Resources like Wikipedia may provide quick definitions where the authority may be adequate for the use.

The interviews revealed that using topic keywords to cast a broad net can be a fairly difficult way for researchers to discover new references outside of their immediate area, although most of the informants questioned said that it is something they do once in a while. One participant expressed this reluctance:

> "I would very rarely attempt to search on 'distributed execution' or something like that. Because you're just never going to get anything. Even in things like academic search sites. I've never really had any good results looking into topic areas that way."

Participants cited a number of different problems when they tried to perform exploratory searches of this sort. Terminology presents one known obstacle, especially if the research is interdisciplinary. Others said that they already have too much to read: "I have a very broad frontier right now. I have a lot of stuff that I can advance incrementally by just following the references." A more senior researcher in his field lamented that he could no longer do exploratory forays into the literature by looking at recent conferences, "There's just too many conferences now."

3.4 Personal Archiving

Earlier I discussed how research materials are stored while papers are being written; in this section, I establish which of these materials are regarded as archival, and how what is archival changes over time. I then describe how materials are stored for long term use and how tools intended for other purposes—for example, email applications or source code versioning systems—have been pressed into service to maintain a personal archive. I examine how changing institutional or professional affiliation is a consistent source of vulnerability for personal archives, trumping many expected problems with formats and media. Finally I explore how participants keep track of authoritative copies of their own work and how they think their archive should be used relative to publishers' or institutional archives.

What is most apparent throughout this discussion is that *personal archiving is a side effect of collaboration and publication*: for example, if email is used as the mechanism for sharing files, it also becomes the nexus for archiving files. If one's CV is the means by which a public list of publications is maintained, it is also used as a pointer for oneself to the most authoritative version of a publication. Personal archiving can be both opportunistic and social: participants talked about tracking down public versions of their own publications to reclaim copies of lost work.

Contents of a personal scholarly archive. What do researchers keep after a paper has been published? How does this view change as time passes? Because most of the participants do not maintain a formal archive, it is difficult to establish exactly what is archival, and what has been left in place in the file system or in email as a side-effect of benign neglect. In other words, as long as a researcher maintains an affiliation with an organization or institution, items may linger simply because the researcher has not bothered to throw them out; it is more unusual for someone to deliberately and methodically cull files than it is to simply declare a paper to be done and move on. While I took care to ask participants about their long-term intentions toward specific items and collections, it is difficult to be sure what they would keep and what they would scuttle, given no real forcing function.

Researchers described at least six types of data that they made special efforts to keep to maintain their intellectual legacy:

- Paper sources and alternate versions of publications;
- The PS or PDFs for the published version;
- Research code;
- Data and logs and the scripts to manipulate them;
- Bibliographies and publications that represent closely related work; and
- Email (individual messages and message attachments).

Of course, this set varies from person to person. For example, some participants do not regard their email as archival; for others, it is the *de facto* substrate of their archive.

Each form of data is problematic to archive for its own reasons. Sometimes code bases and datasets are shared among members of a previous lab; ownership is joint, but an institutional shift generally involves just one person. Or research code may represent intellectual property that belongs to a former employer. Email files may be large and unwieldy—different email applications store messages in different formats. Furthermore, while a researcher may feel that her email is important in aggregate, when she examines it more carefully, she may realize that the constituent messages have different statuses—some belong to a former employer; others are personal; and still others are of dubious long term value and aren't worth the trouble of culling. Data and logs may be large and difficult to re-host. Every obstacle makes it more likely that the item will be left behind.

Over time, all forms of supporting data matter less to participants. Initially participants placed a high value on being able to regenerate a paper from its sources. Some researchers hang onto the versions of the datasets that are used to generate the results (although others do not). Sources for figures and subsets of the data that are used to generate the figures are considered archival too. As time goes by—and a certain amount of inevitable data loss occurs—participants seem to care less strongly that they have "everything."

> "I think I have everything I need. There have been things which other people have asked for, which I would've given them if I could find...People sometimes ask for videos that I used as datasets for old papers. And I typically can't give them those."

How materials are stored to survive. What did participants do when they knew an item was valuable and wanted to ensure it would survive and be findable? Participants have different strategies that have evolved through trial and error. Successful strategies share some common elements, including the ability to:

- bundle related files together;
- establish temporal order and intellectual context (both for provenance and to make items re-findable); and
- be easily maintained, possibly as a side-effect of normal research activities.

Email is cited as a good permanent store for three reasons: (1) it is easy to browse chronologically, which makes retrieval easy and lifts the filing and organizing burden; (2) intrinsic metadata supports the reconstruction of context (for example, who made particular revisions and why); and (3) email is usually accessible from any web browser. If email is used as an archive, some care must be taken to ensure everything that is important is actually in email. Some archival material is normally in email—reviews, for example—and no extra effort needs to be expended to make it part of the record. Other types of artifacts—run output, for example—must be put into email deliberately. Email is a sufficiently good archive that some participants made the effort:

> "I use email extensively as a permanent store of information. … For all of the link ranking stuff I do, I will have run something that computed whatever and I will take the output—just cut and paste it from the command shell and paste it into an email message and mail it to myself."

Zipping up files is another established archiving technique. The compression is not as important as the ability to bundle files into a single unit so no stray pieces get lost. The zip archive may then be named to reflect temporal order and context. Creating a zip archive is also an opportunity to cull and group a set of files. Again, efficient storage is not the objective of this culling—there may be redundancies, for example—but rather the aim is to maintain the associations among (possibly large) groups of files.

Maintenance considerations cause some participants to rely on the file system itself as the archival record. A canonical structure—a "Papers" directory with reliable naming conventions, for example—may be enough to impose order. This method tends to function until the researcher changes jobs.

Finally, some participants use code management repositories such as Source Depot or CVS as a place to deposit archival versions of the papers that go along with the code managed in the repository. The repository is not used to manage paper versions; papers archived this way are not deposited until *after* they are completed.

Of course, none of these strategies totally eliminates entropy and loss: source files become indecipherable as platforms change; files are lost; files that were once intelligible are not intelligible any more. However, loss may not be a bad thing; it may make digital archives more tractable. It is generally acknowledged that after a certain point, it may not be necessary to be able to regenerate publications from their sources. If a publisher maintains a good archive, it may not even be necessary to store the publication itself, especially as older publications become available. One participant said, "I don't care very much [that I can't recreate the electronic versions of my old papers]. I mean, some of those papers are available from the publishers by now." The interesting question becomes, what defines that point? What is the natural curve of entropy? How can benign neglect play out to the researcher's advantage?

One of the greatest hedges against entropy is the index most participants maintain of their own intellectual output. This index takes the form of a Publications web page or, in some cases, a public CV. Not only do these documents keep track of a scholar's intellectual legacy; they also are used to point to authoritative online versions of the documents. Although maintaining this sort of index is regarded as a workable solution, it requires significant effort. Often to fulfill personal and institutional needs a researcher has to edit several different web pages and CV-like documents in the wake of each publication.

Changing organizations as a key vulnerability. Changing organizations is cataclysmic from a personal archiving standpoint. As one participant said, "When you change jobs, you typically lose a lot of things. So my life starts in 2001." To a greater extent than disk crashes or system failures—technological obstacles the participants in this study are well-equipped to overcome—a change in institutional affiliation is responsible for a substantial amount of data loss (this is confirmed by [14]).

Some loss is unintentional: files are misplaced in the shuffle, accounts unexpectedly evaporate, or an organization is becomes defunct. For example, one researcher who had inadvertently lost his publications page said, "[The company] preserved the tech reports, but they didn't preserve the home pages. Which makes sense because it was a defunct organization." Changing organizations may also mean changing platforms and

infrastructure, thus making older files (especially email) more difficult to decode and use.

Other loss is unavoidable: for example, corporate policies dictate that employees must leave all uncleared intellectual property behind when they change jobs.

Recovery from the loss of personal scholarly archives is necessarily partial and usually involves casting about for public copies of one's own stuff; there seems to be some reluctance to ask co-authors for copies of lost work, especially after a significant period of time has passed. Naturally, source files and data files or logs are not recoverable from the public record; these files must either be abandoned, or recovered from co-authors. For example, participant P5, a midcareer researcher said

> "One particular annoyance is that when I changed jobs, I didn't take a big stash of files with me. So essentially anything I did before 2001, which is lots of years of work, is gone, right? And then subsequently I did collect I think pretty much everything from public sources."

In some instances, recovery of one's work from the published record may require a researcher to pay publishers. Furthermore, sometimes what is recovered does not replace what is lost: the recovered paper may have been poorly scanned from print proceedings, or a longer version might not be in the public record. Furthermore, anything that is incomplete (at the time of an affiliation change) is in legal limbo, not in the public domain, and possibly not accessible as licensable intellectual property either.

Changing organizations marks a point in time in which one's digital belongings become a jumble—participants are not quite sure what has survived the move; often digital belongings that are "packed up" are never unpacked again and the archives become inscrutable.

One natural (but dramatic) side-effect of changing organizations is that clever replication schemes—for example, files that have been automatically backed up onto corporate servers or stored on multiple employer-owned computers—are re-centralized; in so doing, they present a single point of failure. Simply put, *change makes digital belongings more vulnerable*. One senior researcher described the effects of a disk crash just as he was moving to a new organization. He hadn't realized that his once well-replicated collection had been reduced to a single copy:

> "I knew that the disk contained basically everything from [my] career… The crash only happened a couple of years ago. It turned out that my laptop was essentially carrying the archive."

4. IMPLICATIONS AND CONCLUSIONS

I began this study with three ultimate purposes in mind: (1) Supporting the collaborative information management (CIM) that is associated with co-authoring a research paper; (2) Supporting the creation of a personal scholarly archive; and (3) Facilitating the downstream deposit of scholarly materials into institutional and disciplinary repositories. I address implications for each.

Implications for CIM. Diverse writing and research styles, coupled with technological heterogeneity, guarantee that there will be varying degrees of buy-in to any CIM infrastructure. Not only that, but researchers may opt in and opt out when the situation demands it, possibly during the course of a single collaboration. Thus—in addition to being usable across different platforms—systems that support information management for cross-institution collaborative writing must:

- Support the abstract notion of a *collection*, bundling together heterogeneous publication files, datasets, and other items (for example, reviews in an email message) that together constitute a research artifact [8]. The items in a collection may be thought of like Fedora's complex objects [10], although it is not clear how this representation handles objects that are at the sub-file level such as email messages;
- Support the designation of a *reference replica*, a copy of the collection that is guaranteed complete, at full fidelity, and up to date;
- Support *filtered synchronization* with collections (so collaborators' local collections can be synchronized subsets of the reference collection), even if co-authors straddle an institutional firewall;
- Support collection *inclusion using metadata surrogates* in the event that the collection includes datasets too large to be copied;
- Support email-like *documentation and chronological organization* of collection elements as they are exchanged. This will index and organize materials in ways that make it easy to find them later (see for example [15]); and
- Support *awareness* in lieu of full synchronization during periods of peripheral participation (e.g. traveling); this may be implemented as metadata-only synchronization.
- Support the *designation of semantically meaningful versions* of collection elements, including datasets and code, maintaining relationships among collection items.

It is easy to see how email provides just enough mechanism to fulfill the minimal version of these requirements. *Any CIM infrastructure must beat email along all of those dimensions if it is to be adopted in email's stead* (see, for example, [18] for a view of how email can be a PIM substrate).

Implications for personal scholarly archives. Earlier I asserted that the construction of personal scholarly archives is necessarily a *side-effect* of sharing, publishing, and backing up files. It is through these normal activities that researchers end up with scholarly archives. This situation seems unlikely to change.

Thus, it is up to us to approach personal scholarly archives with the expectation that they are built automatically and tended with minimal stewardship. While it is seductive to think of these archives as subsumed by the CIM infrastructure specified in the previous section, this is not the long view. Even if archival materials begin their life in another store, it is likely they will need to be maintained separately from that repository, for example, when the scholar changes affiliation. Personal archives must be able to be disentangled from organizational storage (remembering that there are legal issues) and moved. A strategy similar to the one proposed in [13] may be appropriate.

What is the role of bibliographic resources in a personal archive? Although researchers had immediate reasons for storing their bib files the way they did, conceptually a master bibliography is important. Along with a bibliography, foundational and central

references may be kept, along with annotations and comments that reflect particular readings of these references. *Along with a scholar's own work, a personal scholarly archive should provide a place to store one's personal digital library*.

Other research artifacts—code, datasets, simulators, reference corpora (that is, standardized datasets)—play a significant role in the work of these computer scientists. But are they archival? Right now, they aren't. Many participants confessed that they could not regenerate their published results because they had not archived intermediate datasets, datasets that were dependent on network state and other circumstantial factors (compiler parameters, e.g.). Is it possible to save these all of these artifacts? Is it necessary? This is something that must be determined by the scholars' research community; the ability to reuse the data fundamentally changes the nature of the science [12].

Implications for institutional and disciplinary repositories. The most important implications of this study for institutional repositories stem from notions of *agency*: what human actions are necessary to deposit research artifacts—publications, datasets, simulations, and code)—in institutional repositories, especially considering that it is likely that more than one author of a given publication must act. As I noted earlier, different authors have different responsibilities; the author that has the final version of the paper in hand may not even have *access* to the final version of the dataset. To make matters more complicated, the dataset and code used in the evaluation phase of the research may be 'owned' by a different member of the collaboration than the system code. Furthermore, in several of the study's sub-disciplines, datasets exist in many intermediate forms, all of which are meaningful, some of which are private, and none of which may be fully documented. *If nothing else, this study highlights the need to consider agency when we talk about depositing new units of scholarly communication into disciplinary repositories*.

What is striking is how much overhead is inherent in maintaining our current scholarly communication systems. It is not unusual for a researcher to have to supply relevant files and bibliographic metadata to multiple different places once a paper is published. The files are not only sent to the publisher, but also may be required by an institutional repository and other local (and possibly competing) stores. Each co-author is also maintaining her own Web page, updating her CV, and trying to ensure she has the most recent versions of all the files used by the publication. It's no wonder that ones' own files are maintained opportunistically.

Can this overhead be reduced? Can the metadata be enriched? At many points, we know more about the research artifacts than we do at deposit time. For example, one participant described stripping the headings from of his dataset so he could feed it into MATLAB. This problem is exacerbated in the case of papers that are published well after they are written, either because they are rejected from their original intended venue or because of an extended review cycle or both.

It is easy to see ways in which collections may be archived as a side effect of supporting the collaborative production of scholarly artifacts—at the point at which the files are shared, we have them in our grasp—but it is more difficult to understand how repository quality can be improved without adding a great deal of unwanted overhead to an already onerous process. From a scholar's perspective, the challenges raised by personal, institutional, and disciplinary repositories are many and they are far from solved.

5. ACKNOWLEDGMENTS

Thanks to all of the researchers who gave so freely of their time for this study. Thanks too to my adopted group at MSR-SVC, Doug Terry, Ted Wobber, Rama, Tom Rodeheffer, and our trusty intern Meg Walraed-Sullivan.

6. REFERENCES

[1] Adams, G. 2006. Beyond OAIS. *Proc. Archiving'06*. (Ottawa, Canada, May 23-26, 2006), p. 7.

[2] Bolter, J. 1991. *Writing Space*. Hillsdale, N.J.: Earlbaum

[3] Borgman, C., Wallis, J.C., Enyedy, N. 2007. Little Science Confronts the Data Deluge. *IJDL 7*(1) 17-30.

[4] Bowker, G. C. 2000. Work and information practices in the sciences of biodiversity. *Proc. VLDB 2000,* San Francisco: Morgan Kaufmann, 693-696.

[5] Diaper, D. 1993. Small-Scale Collaborative Writing Using Electronic Mail. *CSCW in Practice*, Germany: Springer-Verlag, pp.72-102.

[6] Foster, N. F. & Gibbons, S. 2005. Understanding Faculty to Improve Content Recruitment for Institutional Repositories. *D-Lib Magazine, 11*(1).

[7] Hey, T. and Hey, J. 2006. e-Science and its implications for the library community. *Library Hi Tech, 24*(4), pp. 515-528

[8] Hitchcock, S., Brody, T., Hey, J. and Carr, L. 2005. Preservation for institutional repositories. *PV2005*, The Royal Society, Edinburgh, Scotland.

[9] Kintsch, W. and van Dijk, T. 1978. Toward a Model of Text Comprehension and Production. *Psych Review* 85, 363-394.

[10] Lagoze, C., Payette, S., Shin, E., and Wilper, C. 2005. Fedora: An Architecture for Complex Objects and their Relationships. *IJDL*. http://arxiv.org/abs/cs.DL/0501012.

[11] Lynch, C. 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report*, No. 226, pp. 1-7.

[12] Lynch, C. 2007. The Shape of the Scientific Article in the Developing Cyberinfrastructure. *CTWatch Quarterly*, August 2007.

[13] Marshall, C.C. 2008. Rethinking Personal Digital Archiving. *D-Lib Magazine*, *14* (3/4), 2008.

[14] Marshall, C.C., McCown, F., and Nelson, M. 2007. Evaluating Personal Archiving Strategies for Internet-based Information. *Proc. Archiving '07*. Springfield, VA: Society for Imaging Science and Technology, pp. 151-156.

[15] Ringel, M., Cutrell, E., Dumais, S., Horvitz, E. 2003. Milestones in time. *Proc. Interact 2003*, pp. 228–235.

[16] Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., and Warner, S. 2004. Rethinking Scholarly Communication. *D-Lib 10,* 9.

[17] van Ingen, C. personal communication.

[18] Whittaker, S., Bellotti, V., Gwizdka, J. 2007. Email as personal information management. *CACM, 49* (1), 68-73.