from writing and analysis to the repository: 
taking the scholars’ perspective 
on scholarly archiving

Cathy Marshall
Microsoft Research, Silicon Valley

JCDL 2008, Pittsburgh PA
18 June 2008
research questions

• How can we support the deposit of research material into scholarly archives?
• When we say research material, what do we mean?
• What goes into a personal scholarly archive and how can we make this stuff less vulnerable?

look upstream for answers...
Building institutional repositories is harder than it looks. Viruses in deposited files (e.g. Adams, 2006) motivation for deposit is sometimes absent (e.g. Foster and Gibbons, 2005) datasets might not be well documented (e.g. Borgman et al., 2007; Bowker, 2000)

There are good reasons to share more than publications. Sharing may be outside existing practice (e.g. Borgman et al.) sharing may support new types of scientific results (e.g. van Ingen)

Scholars (seem to) want to archive stuff for themselves
study

15 industrial CS researchers; 16 interviews; 6 months of participant observation

- semi-structured, open-ended interviews
- photographed, recorded, transcribed, analyzed

- interviews ranged from 45 to 90 minutes
- interviews were artifact-centered

- participants were researchers in different phases of their careers
- publication is an important part of what they do
### CS subdisciplines represented in the study

<table>
<thead>
<tr>
<th>subdiscipline</th>
<th># of participants*</th>
</tr>
</thead>
<tbody>
<tr>
<td>algorithms and theory</td>
<td>3</td>
</tr>
<tr>
<td>distributed systems</td>
<td>11</td>
</tr>
<tr>
<td>security and privacy</td>
<td>5</td>
</tr>
<tr>
<td>software tools</td>
<td>3</td>
</tr>
<tr>
<td>web search and data mining</td>
<td>4</td>
</tr>
</tbody>
</table>

*many participants identified with two (or more) subdisciplines*
artifacts/subdisciplines
representative theory paper

- proofs; equations; theorems and lemmas
- often no datasets
- often no figures
- produced by smaller group collaborations
• often contain ‘drawn’ diagrams created in graphics editor

• small to medium-sized datasets may be used for evaluation

• analysis software (e.g. Excel or MATLAB) may be used to generate some figures.

• produced by larger group collaborations (4-7 people)
may be illustrated with graphs and charts; some of the data may be generated by the research software;

• mathematical expressions

• very large datasets; may involve several types of datasets

• produced by small-medium collaborations (2-4)
• may have pseudocode snippets as figures

• may have small datasets (performance evaluation)

• produced by varying sized collaborations (tending toward small groups)
findings
findings in 4 key areas

- collaboration and writing
  setting the stage: people, devices, and technology infrastructure; upstream activities; writing roles

- storing and managing research material
  what kinds of stuff are we talking about? how is it organized? what’s archival and what’s not? Thorny issues re: datasets...

- finding, managing, and sharing bibliographic resources
  how a web of related work plays into maintaining a personal archive; how this manifests itself in artifacts

- personal archiving
  what’s actually part of a researcher’s intellectual legacy? how are materials stored to survive? how do people lose stuff?
for each finding I’ll...

- break it down
- show you a few examples
- talk about why we care
collaboration and writing

roles in writing

places and devices

infrastructure and tools
different degrees of involvement
• different skills and backgrounds
• different infrastructures
• different resources
• different institutional affiliations
• different work habits

this heterogeneity shouldn’t be a surprise, right?
roles in writing: some examples

“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.”

“Some of the figures were initially done by Mike* in PowerPoint and Paul did the other figures in Illustrator. ... The graphs I forget who generated them. ... For the final copy... Paul is not a TeX person and so he generated EPS for me and I actually made them appear in the text with the formatting and partial columns and all that kind of stuff. Because I'm familiar with TeX macros.”

*Names have been changed
roles in writing: why do we care?

- each author may end up with different pieces that contribute to the final product
  e.g. the collaborators evaluating system performance may be the only ones with the datasets

- some artifacts may not be uniformly meaningful across the collaboration
  e.g. some collaborators will not know what the individual fields in the datasets mean

- some artifacts may not be reproduced, edited, or viewed by some collaborators
  e.g. some collaborators might not have the MATLAB license necessary to produce the figures
places and devices: break it down

- no need to take mobile computing too literally: people relegate different activities to different places and devices (remember what Bill Buxton said yesterday: place matters!)

- pragmatic constraints dictate what activities are associated with particular devices
there are always exceptions to the rule
“[I worked on] everything [having to do with the paper] here. I cannot focus at home... I have three kids and the noise level is such that my brain gets randomized within fifteen seconds...”

“I find that paper writing is most effectively done here in this chair with a large enough display in front of me.”

“I commute by train from San Francisco. So when I'm writing a paper, I often edit text on the train on my laptop. And to a lesser extent, at home. But when I'm in the office, I'd be using my desktop. And I just copy things manually back and forth between them...”

“The last [final] version was on my other computer and I haven't moved it onto this computer yet.”
places and devices: why do we care?

- it is likely no set of files on a single device represents ‘ground truth’ for the publication
  
  the ‘penultimate effect’: authors might not synchronize all their devices once the paper has been submitted

- sometimes legal, security, storage, bandwidth, or cognitive limitations enforce incompleteness
  
  all files related to a publication may not be replicated everywhere
Although participants frequently alluded to homogeneity (“We live in a LaTeX world”), there are lots of supporting applications (and a few secret Word users)

(At least) 4 categories of paper-writing tools:

- content editors (in addition to LaTeX/emacs)
  * e.g. Visio; Illustrator; Powerpoint; Xfig; Word
- data analysis tools
  * e.g. Excel; MATLAB; GnuPlot
- infrastructure software
  * e.g. Outlook; Source Depot; CVS
- custom software (apart from research sw)
  * e.g. scripts to strip off column headings for MATLAB
“So I think a lot of the high-level work was done on a pad of paper. But I usually don't believe a result until it's written up in LaTeX... Usually what happens is, you have a sketch of the result on paper, and you know it, before you start writing. Then you start writing it and you realize that there are some things that you didn't think out clearly that you have to fix... The LaTeX's usually not sketches. Usually someone's gone and written every single detail. It doesn't leave pieces to the imagination.”
Heterogeneity makes complete preservation of research materials costly—it becomes important to know what to keep and why to keep it.

- canonicalizing formats v. only saving sources v. saving all originals

Any kind of upstream infrastructure to support writing will need to be flexible and general

- email is the lowest common denominator—and most effective—support for collaboration to date
storing and managing materials

managing related material

managing versions

managing data
managing related material: break it down

• publications may *drive* the archive, or publications may *organize* the archive
  *in other words, the archive may be the paper + what it took to produce the results and nothing more or... the papers may be simply the organizing force for all kinds of related material*

• other (possibly) important factors
  *completeness*—everything necessary
  *up-to-datedness*—final version of everything
  *tidiness*—eliminate intermediate artifacts
  *versions*—versions are kept and their semantics recorded
“So here's the directory that houses that particular paper. And it includes figures from the paper, PowerPoint presentations related to the paper. You may also see that it contains some things that have little to do with the paper.”
managing related material: why do we care?

- gradual transition from a warm store to a cold store

  the contents of a personal scholarly archive may change over time (how long since they’ve been active?)

  intermediate versions, supporting material, and related files fall into the bit bucket first

  As time passes, researchers feel less of a need to reproduce the publication from its constituent parts—the final PDF/PS file is all they need
managing versions: break it down

• participants pine for (and sometimes use) versioning systems. Why?
  a worry about managing conflicts and coordinating changes
  as backups to stave off catastrophe
  to recover clever turns of phrase

• but there are less immediate reasons too
  meaningful checkpoints in the research lifecycle that provide stable points of reference
  a means of recording the history of a set of ideas
managing versions: examples

“This is a conference version. ...I'm making it into a journal submission... [which means that] basically we rewrote it. We improved every single result and every single section. ... The approximation was four in this draft. And we improved it to three. Which is a better constant.”

“I can ... give you the version of that file at any particular date you want. This is my own VC system. It's some emacs macros I wrote, so that whenever I save a file, it does an RCS version. So. Since I grew up in the days that computers crashed every half hour, I save files often.”

“I guess I've done this on occasion, which is to keep a copy often just a PDF version when I've made some changes so that if I get a question, someone asks you know, "What about this thing on page 37?”, I'll be able to find it because I'll have the same [pagination].”
Participants are computer scientists. This must be taken into account.

Actually two different questions

Are versioning systems necessary?

Are versions necessary (in the long term)?

Email, naming conventions, and social arrangements may get you far enough and may have advantages

In the long term, versions are only useful if the version is a separate publication
>2/3 of the participants maintained significant datasets

- datasets vary greatly (from a small number of points in a graph to terabytes)
- some are impossible to recreate, but are straightforward to generate
- some (e.g. machine learning training sets) are costly to construct
- some datasets drive writing; others are just illustrative

spectrum of management practices

- few are archived or even backed up
- some have their own hardware; a few are just stuffed into the LaTeX file, as is
- many follow a complex trajectory and are corrected, gap-filled, or post-processed
managing data: examples

“It takes a really long time to move terabytes of data...the link from here to there ... is only a gigabit anyway. And then the link from the lab to the building is only a gigabit link and from the building to the rest of the campus is only a gigabit. It's just easier to get off the train in Palo Alto, walk over to [the data supplier], pull up—port literally 4 terabits of disk in a little cardboard box, come back here and plug them directly into the machines.”

“I just put the performance data inside the paper source file as comments. Just so they'd always be there if I needed them.”

“This is a case [fixing input data for a simulator] 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.”
managing data: why do we care?

• Backdrop of re-use and scientific integrity: which datasets do we save?
  The value of the datasets varies greatly
  Make the most of labor-intensive datasets

• When in the analytic arc do we save them?
  raw or manipulated? (e.g. ‘fixed’)
  complete or excerpted?
  code or derived data (e.g. the crawler or the crawled?)

• How do we store and document them?
  what about associated code?
  what about processing elements like input parameters or compiler options
  headings may interfere with processing
managing bibliographic resources

maintaining local resources

sharing references

non-traditional resources
maintaining local resources: break it down

- BibTeX files as a cumulative investment
  - centralized v. template for pub topic

- intellectual bookkeeping
  - what researchers have read
  - what they want to read in the near future
  - where they found it
  - maintaining local copies of published material

- distinct types of resources
  - central
  - foundational
  - peripheral
maintaining local resources: examples

“Actually there's a few papers in that bib file which I [haven’t read]. The ones which don't have comments, it means I didn't read them yet. So it's "to do"... if I don't plan to read them, I don't put them in there.”

“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.”

“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.”
Bibliography is important in its own right

Bibliographies may be personal and personalized

Bibliography = view onto personal digital library = view of scientific community membership and intellectual allegiances
sharing references: break it down

• a comprehensive bib file is an intellectual dowry

  related work specialists

• public/private split for shared bibliographies

  the entry might be sharable, but not the comments
sharing references: examples

“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... Because I can't translate my terminology into their terminology.”

“Comments are very sensitive you know. Sometimes you say, 'I didn't like this paper.' ... Once you give them [bib files] 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.”

“Mike* wrote the related work section and he has an incredibly beautiful bib file. So he actually dealt with that for this paper... I've never seen anyone with as good bib file as Mike.”

“Well, you know, the references are useful if you've read them, but what's the point of having somebody else's references?”
sharing references: why do we care?

- may be a way of taking advantage of others’ background and disciplinary knowledge. (c.f. Zotero; CiteULike; etc.)

- an important aspect of collaboration—bringing recommenders into scholarship
non-traditional resources: break it down

- Increasing use of non-traditional references like blogs & wikipedia
  - *good window onto the doings of expert practitioners*
  - *good source for definitions*
  - *but concern about stability and authority*

- General search engines (i.e. Google or Live) are used as starting point rather than Google Scholar or Live Academic
  - *utility tied to coverage, scope, authority, and timeliness*
  - *general search engine used as access point for DL material*
  - *reverse search locates ‘authoritative’ copy*
non-traditional resources: examples

“Let me give you just one example. This guy has a fantastic blog... and he writes about a lot of issues in data centers. So I do data center computing. ... But now this is a completely new area for me... There's a lot of links and PowerPoint presentations and stuff. And he blogs almost every day.”

“Probably the thing that I usually need to do is find definitions of things. I usually try Wikipedia or something like that.”

“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.”
non-traditional resources: why do we care?

- compensating for unstable nature of some of these resources
- citation issues suggest a need for stability
- ways of fostering awareness of non-traditional resources need to be integrated with ways of fostering awareness of more traditional ones.
personal scholarly archiving

contents of a personal archive

vulnerabilities
personal archive contents: break it down

• what researchers keep to maintain their intellectual legacy
  • *paper sources and alternate versions*
  • *PS/PDFs for published version*
  • research code;
  • data and logs and scripts
  • bibliographies and related work
  • email (e.g. reviewer comments)

• every form is fraught for its own reason
  • shared ownership (cross-institutional)
  • intellectual property
  • unwieldiness of large data files
  • material that’s valuable in the aggregate, but not individually (think email)

• CV acts as an index

*picture of Esther Dyson’s office courtesy of Ramana Rao’s blog*
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.

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.

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.

There is some information on which is the best version of a piece of work. On my CV. So when you were asking me which is which for each of these files, I can't tell you what each of the files is, but I could find out which is the best version in some way.
personal archive contents: why do we care?

- Personal archiving is a side effect of collaboration and publication: we’re most likely to end up keeping what we share and publish

- It’s easy to observe the effects of benign neglect—should we be fighting them? Or should we focus on what we’ve demonstrated need for?

- It’s easier to keep than to cull, but it’s easier to lose than maintain.

- A cv (or a web publication list) may be the best index
vulnerabilities: break it down

- The key vulnerability is *changing organizations*; it is more cataclysmic than technology failures.

- Loss may be unintentional
  - files are misplaced in the shuffle
  - accounts evaporate more suddenly than expected
  - organizations become defunct
  - infrastructure changes
  - digital belongings become a jumble
  - replication schemes are re-centralized

- Loss may be unavoidable (e.g. uncleared IP)
vulnerabilities: examples

“When you change jobs, you typically lose a lot of things. So my life starts in 2001.”

“[The company] preserved the tech reports, but they didn't preserve the home pages. Which makes sense because it was a defunct organization.”

“One particular annoyance is that when I changed jobs, I didn't take a big stash of files with me. ... And then subsequently I did collect I think pretty much everything from public sources.”

“There are a couple of things that I cleared [from my old company] and took. But I didn't bother to clear everything and I didn't take anything I hadn't cleared. There was in particular one piece of research that I was working on that I hadn't yet written up—and I cleared that and took it. And then wrote a paper and it didn't get accepted. And so it's actually now in this weird limbo because everybody would've been much happier if it had been published.”
vulnerabilities: why do we care?

- is there a need for a personal scholarly archive apart from disciplinary or institutional archives?

- In the end, changing affiliations may be the best way of cleaning house.
implications
Implication 1: any upstream collaborative infrastructure needs to beat email!

email maintains a version chronology
changes are documented
informal locking mechanism
heterogeneous information is brought together (e.g. reviews, discussion, conference information)
exchange is flexible (i.e. not all authors get all versions)
firewalls don’t pose a problem
files can be bundled with zip etc.

*Example taken from my own email; not study data*
Implication 2: Personal scholarly archives are not just the file system or email folders + 20 years

- much of what we keep today does not seem to have long term value; what remains to be investigated are potentially reusable bits, in particular datasets and bibliographic resources

- whatever we keep must be disentangled from the institutional storage
implication 3: we need to take advantage of upstream practices in downstream deposit and archiving

- Agency, agency, agency: the scholar who deposits the paper may not be the scholar who deposits the dataset.

- There is tremendous overhead for scholars in maintaining our current scholarly communication systems. Files and metadata often must be supplied to multiple places, and similar updates must be made to local documents and web pages.

- We may know more about our research artifacts (e.g. datasets) at other times than when they are actually deposited.
“Some papers are written in a very short time, while others take many years and move across organizations and continents.”
questions?

http://www.csdl.tamu.edu/~marshall
http://research.microsoft.com/~cathymar