The future is data, not text

We’ve largely finished the out-of-copyright books.

The next step is data: there is vastly more scientific data than there are creative works.

The data needs more attention than books do – we have less experience with long term storage and organization.

The problems are more social than technical or economic; we do not reward the activity of keeping and using old materials, and so it’s not done well.
Reviewing: books and their users

Do readers want paper or pixels?
Books are not really paper

“There are two meanings of the word “book” ... there is one meaning which signifies a physical object ... another meaning signifies an intellectual object ... the physical book is never more than an ingenious and often beautiful cipher by which the intellectual book is communicated from one mind to another ... if it is not the physical book but the intellectual book of which the librarian is keeper, then his profession ... is not the profession of the custodian, for the intellectual book is not a ticketed parcel which can be preserved by keeping it from mice and mildew on a shelf.”

Archibald MacLeish, *A Time to Speak*, 1941.
Don’t fixate on Google, it’s about Amazon and Barnes & Noble

Each year in the US:
about 3 billion books are purchased.
about 2 billion books are borrowed from public libraries;
    nearly all are recently published, in-print books
fewer than 40 million books are borrowed from research libraries.

So almost all reading is of in-print books, which is not the focus of Google library scanning.

Amazon has sold 1M Kindles, and for books available in Kindle format, Kindle sales are 35% of print sales.
Too much duplicated scanning: 7 copies of each book!!

I took 13 more or less random out of copyright titles: all books in the Rutgers library with either penguin or Paraguay in their title, published 1860-1920. The only one you might have heard of is Anatole France, Penguin Island.

Google has scanned these books 68 times, and the Open Content Alliance has scanned them 30 times. Despite this, there are two titles that wouldn’t be in Open Content if they weren’t picking up all the Google scans they could. One of these books is not available as “full view” in either system (although the copy Google scanned was published in 1903 and the author died in 1906).

So it’s time to move on from 19th century books.
Students using online books

Students asked to choose books for a paper did about as well whether they had the book on paper, or on PDF. They also took about the same amount of time (although this was influenced by the experiment).

However, the students who did best were most able to use the online books; the students who did not do well performed better with paper.

They browsed in substantially different ways: with paper books, students went to the table of contents and the index. With PDF, everybody searched, despite problems designed to make searching difficult.
Where readers look

Figure 1. Political History of England, paper
Figure 2. Political History of England, PDF
The scientific paradigm changes

From:
  invent hypothesis, think of data needed to test it, design experiment, run experiment, evaluate hypothesis, ... and repeat for a career

To:
  invent hypothesis, think of data needed to test it, look up data in an online databank, evaluate hypothesis, ... and repeat faster

Data moves from “just in time” to “just in case”

But how do we know the old data will be there for us?
Research Data & Information

Among the issues in asking for data to be public:

Private, time-embargoed, or public?
How encourage preservation and use?
What are the implications for privacy and national security?
How preserve data quality under open access?

What government policy will create the greatest social good?

Who should work on saving it? The Internet Archive doesn’t use any subject specialists; libraries and museums traditionally did. Can you save astronomy data if you are not an astronomer?
What should we save?

Bill Arms once said that we would be dividing our information into three piles: (1) valuable enough to justify manual effort to preserve; (2) worthless enough to throw out; (3) stuff that would be kept as a pile of bits in the hope that our successors will be have better automated tools to deal with it. And he suggested that the last pile would be 90% of the whole.

Disk gets cheaper: anything you can capture you can keep.

Searching replaces metadata.

Society uses more information, so we’ll keep more. But it won’t be read by people; it will be scanned by machines.
The history of disk prices

Will cloud computing change this?
Too much data even for all this disk

From “The Diverse and Exploding Digital Universe” by EMC.
Attention is the scarce resource

The future is
  skimming & searching
  snippets
  multimedia
  long tail

In scientific journal reading, online is already dominant.

More books are available to more people; in all but 11 states, Google Book Search is the biggest library in the state.

Scholarship used to be a cat sitting patiently at a mousehole; it will become vultures circling the countryside.
Life Before Google

How can he bury a bone today, and then find it three months from now, especially under all this snow?!

I just don't understand.

It's very simple.

I have "total recall"!
Searching

Several problems that seemed intractable for decades are now becoming practical:
  voice searching (Google on Android)
  face searching (Picasa, Fuji, Riya, ...)
  landmark detection (Google, Yahoo)
We’re still working on
  handwriting recognition
  3-D searching
  objects without sharp corners
  data base creation

But there’s hope for almost anything we can save. If we just keep it, the future can probably find it.
Face search (Google Picasa)
From right to left, Google from above, Microsoft at 45 degrees, and Amazon at street level; Google now provides the street level view.
Bridgeport: aerial history

This industrial portion of a once-vibrant landscape is now starting to see single-family homes again.
Data not intended for humans

Sensors are producing petabytes of everything: medical imagery and data, video surveillance, scientific data from experiments, …

Molecular biology and astronomy are the first online, shared-data, sciences; but lots more is following.

The whole style of collecting data will change: it will stop being just-in-time and go to just-in-case.
Sky pictures
Astrophysics Data Service

This stores over 300,000 articles (and 140 TB or so of numerical and image data).

There are about 12,000 astronomers in the world; they read about 1.2M journal articles on paper, and about 3.4M online. Michael Kurtz estimates that time saved using the ADS is equivalent to an additional 736 astronomical researchers, or a monetary gain of $240-$300M.

In the top institutions, it is now unusual to see somebody reading astronomy journals on paper.

The journal articles are about 500 GB, so they are a trivial fraction (0.4%) of the data size (and easier).
Protein Data Bank

Nearly 45,000 deposited structures;
in the chart, the orange bars plot total structures, the blue bars yearly increments.

(Helen Berman, Rutgers; also UCSD)
IRIS seismic consortium

40 terabytes in 2004; growing 4 TB/year (U. Washington, lead institution; Tim Ahern, many others.)
Long term data storage: It's a people problem

People need to be trained to save data, funded to do it, and rewarded for it.

Whom do you promote? The guy who writes a new program or the guy who finds that it's already written?
“One small step for a man”

... LOST

“Houston, our tapes have gone missing

Houston, we have a problem. There is probably no artifact in the history of space exploration more precious than the first television images of the Moon captured by Neil Armstrong and his fellow astronauts as they disembarked from their lunar module in July 1969.

Unfortunately, the magnetic tapes of those images have gone missing. Worse still, they appear to have been missing for at least 30 years - and nobody, until now, even noticed.”

(The Independent, Houston, August 13, 2006, and many other newspapers.)
Not the first time, unbelievably...

“Lost Moon-landing tape found

Tense moments before the touchdown
By BBC News Online science editor Dr David Whitehouse
A dramatic recording of the first manned landing on the Moon has been rediscovered at NASA's Johnson Space Center in Texas.

The tape covers the crucial few minutes as the Apollo 11 lander touches down on the surface of the Earth's satellite in July 1969.

... It was found in the audio library at NASA's space centre in Houston. The recording had been labelled "bad tape" because it was in a very poor condition."

(BBC, Sept. 19, 2001; this is an audio recording and is not the video referred to in the previous slide).
Back to NASA

The tapes of the moon landing that are lost were in an unusual format (remember in 1969 we didn't have DVDs or even VHS cartridges) "The only known equipment on which the original analogue tapes can be decoded is at a Goddard centre set to close in October, raising fears that even if they are found before they deteriorate, copying them may be impossible" - Sydney Morning Herald (and other papers).

Format incompatibility, in a world where we change formats constantly, is a major issue.
Keeping the original data

When Sir Alexander Fleming published his original paper on the anti-bacterial effect of the mold Penicillium, he misidentified the exact species; he reported it as *Penicillium rubrum* instead of *Penicillium notatum*. If somebody at Oxford hadn't kept the actual mold around, the wrong species would have been tested, found to be almost inactive, and the work abandoned.
The social problem is the hardest

“A general problem prevalent among all scientific disciplines is the low priority attached to data management and preservation by most agencies. Experience indicates that new research projects tend to get much more attention than the handling of data from old ones, even though the payoff from optimal utilization of existing data may be greater.” - Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources (National Academies Press, 1995)
Even clay tablets aren't permanent

“The burning of the Persian palace-site of Persepolis after its fall to Alexander, for example, although a savage act of vandalism, contributed to the preservation of the palace archives. In the 1930s excavators recovered this archive inscribed on unfired clay tablets. Under most conditions clay tablets are, by their nature, more durable than other types of media. The Persepolis tablets were written to track economic transactions. The scribes who recorded them would perhaps have been surprised that by analysing these thousands of tablets, it proved feasible to profile the position and role of women in ancient Persia under Darius I to Artaxerxes III. Sadly, only a percentage of these tablets have been fired since their discovery in the 1930s and many are reported to be drying out and crumbling away in their new home at the Oriental Institute at the University of Chicago. The content of many of the tablets has not yet been transcribed and mere recovery of media does not necessarily protect it or its contents against loss.“ (Seamus Ross, Changing Trains at Wigan, NPO, 2000).
Can we just give the problem to the libraries?

As a professor in a library school, I wish I could say that libraries were the obvious organization to take care of data. They understand keeping things for a long time and arranging to find them later. It would be a sensible new activity to balance a decrease in foot traffic into book collections. But...

• They have not been ambitious in this area; libraries feel under budget pressure and don't want new tasks.

• They lack the subject area knowledge to deal with complex data sets in scientific areas

• They often lack the technical skills for advanced data handling.
It’s not about buildings any more.

Users are increasingly online but libraries still define themselves as physical buildings.

A central campus library with architecture:
• Lewis Science Library, Princeton: $200/book
• Seattle Public Library, $130/book

Without architecture
• Gallagher Law Library, U Washington: $60/book
• Cal State San Marcos: also $60/book

Off-campus warehouse: $1-$3 per book.
Libraries

Seattle history

1906: $2.50/bk, $4/sq ft
1960: $4.50/bk, $22/sq ft
2004: $114/bk, $506/sq ft

(not inflation adjusted)
Making data useful and usable

The British Library wrote in its strategic plan some years ago that it did preservation for its future users and access for its current users. Only current users, however, vote on today's budget. Data curation is not likely to be supported unless it also improves current access.

Fortunately, this goes with long term needs: many of the same steps are needed to make the data easily used as to make it easy to preserve. These include translation to standard formats, encoding of metadata, and recording all necessary rights permissions. And this is easier to do at the time rather than retrospectively. Some other steps such as visualization and statistical software for the data are keyed to current users, however, and are of less importance to the future.
Who should save what?

The US is moving to university-based data archives; the UK has had discipline-based systems. Other organizations play significant roles: one notes the San Diego Supercomputer Center (Sid Karin recognized years ago that although centralized computing might be decreasing in importance, centralized data was growing).

We also have organizations such as NARA, LoC, the Internet Archive, and university consortia such as ICPSR (social sciences) and IRIS (seismology).

The UK has a “Data Curation Centre” for shared expertise; is this a model to copy?
Volunteers

A surprising amount of work is done by volunteer communities: look at Gutenberg, “distributed proofreaders”, Wikipedia, SourceForge, …

Historically, a great deal of data collection and curation was done by volunteers; look in any natural history museum. Today, look at all the Google Earth mashups.

We are starting to see amateur metadata; whether this will be useful in more data-oriented material is not yet clear.

(And in any case, the quantity of material being saved will overwhelm even volunteer taggers).
Traditional archivists

Skills such as paleography (reading of old handwriting) have obvious analogies in the digital age. So do skills such as understanding materials conservation, the organization of materials in large groups to avoid excessive cataloging expense, and dealing with a variety of media. But most important may be the social milieu in which librarians and archivists were traditionally open and sharing of their materials and their knowledge; we should not require each scientific community to develop its own expertise in database preservation. On the other hand, how much subject expertise is needed by the data archivists?
Economics 1: Landsat data

Landsat 1 was launched in 1972, and NASA or NOAA sold the images for relatively low amounts, $400-600. In 1983 the Reagan administration decided to privatize Landsat images (by this time we were at Landsat 5, flown in 1984) and Hughes created EOSAT to sell the images. Prices went up to $4000 or so. As a result relatively few were sold and large areas of the earth had no saved imagery.

In 1986 the French SPOT satellite flew, and prices dropped a bit ($2000-3000). In 1992, with EOSAT unable to make a profit and with Landsat 6 having failed at launch, the privatization was reversed and prices dropped. Other satellites were now up there (Quickbird, Earlybird).

In 2008 the historical Landsat archive was made available free.
Mt. St. Helens, left 1973, right 1983. (The eruption was May 18, 1980).
Economics 2: Weather

Peter Weiss of NOAA (Borders in Cyberspace) writes that commercial meteorology is measured as follows:

US
$400M-$700M, 400 firms, 4000 employees

EU
$30M-$50M, 30 firms, 300 employees

[these numbers look suspicious to me]

You can buy 15GB of US historical weather from NCDC for $4290. Germany charges 4000 DM for one station; one country charges $1.5M for all their data.
Cost recovery doesn't really work

UK Met office: revenue from data sales not significant

British Ordnance Survey: 32% of revenue from private purchases, all the rest government-mandated purchases

Deutsche Wetterndienst: 1% of costs covered by data sales.

Peter Weiss (*Borders in Cyberspace*) complains that data is being withheld from international weather archives in the hope of getting paid for it. There is now an international agreement on at least basic observations being public.
Europe compared with US

In general, US agencies are more likely to make data available at distribution cost; EU agencies are more likely to charge.

The US market for public information is 2-5 times as large as the EU market, despite the EU being about the same size (in 2005).

According to the EU in 2000

<table>
<thead>
<tr>
<th></th>
<th>EU</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investment value of public info</td>
<td>9.5</td>
<td>19</td>
</tr>
<tr>
<td>Economic value</td>
<td>68</td>
<td>750</td>
</tr>
</tbody>
</table>
Data quality

Can we have open and easy distribution with adequate quality? Ian Irvine (then chair of Elsevier) said in 1996 that the papers rejected by his journals were what you found on the web.

To quote James Boyle (Duke Law School), we need to reconcile two true statements: most of what is on the Internet is either incomplete, inaccurate, badly written, or in some other way defective; but you haven't opened a paper encyclopedia in ten years.

Do people care? Online maps have many inaccuracies, and if you go to a cartographic conference you'll hear about them, but web mapping is now dominant.
What we would like to happen

We need best-practices advice to the data centers, and strategic advice for the policy makers.

Astronomy has no commercial value. But the world is engaged in a copyright-extension competition (Mexico leads, life + 100) and the Chinese have been very secretive about their seismology data.

We need a solid evaluation of the impact on science. How can we measure the impact on astronomers or molecular biologists, the two most advanced fields, of public data?

Should there be an “information commons”? Suggestions welcome.
Data sharing ethics

Vary by field

• Molecular biology: you can’t publish a paper reporting a protein structure without depositing the structure in the public data bank. Genomic data also public.

• Astronomy: convention is you get two years’ use of the data you collect, then must make available to others

• Dead Sea Scrolls: kept secret for forty years.

Yet molecular biology data has potentially enormous economic value, whereas cosmology and ancient scrolls have none.

What should we urge on new fields?
Conclusions

We're seeing everything go online

- First it's said that it's just a finding aid.
- Then it becomes the material for human research
- And then it becomes used for data mining

Technology problems: searching sounds or images

Economic problems: funding curation

Legal problems: mostly copyright

Social problems: encouraging cooperation and new skills