

# Research Data & Information

Science is fundamentally changing: data collection is being separated from hypothesis formulation and evaluation.

The National Academies have a new board on research data and information, to make suggestions for data policy. The goal is to inform and advise government agencies about how they should treat scientific data and publications:

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?

# Open access publication

Last year's battle; but what should be done about the research data on which the papers are based?

If data is available, other scientists can  
validate it  
do meta-analysis  
combine with data from other disciplines

Web mashups are an example of the benefits of combining data from different sources.

# 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?

# Astronomers

Perhaps the most successful group has been astronomers. There are only about 10,000 astronomers in the world, and they share their data through the Virtual Observatory projects. There are standard data formats and access methods, plus massive databases.

A good deal of this is due to an individual – Jim Gray – who carried through the design of the Virtual Observatory and the databases, partly from personal interest and partly persuading Microsoft to support astronomy as an example of an enormous database.

The other equally important leader of the project is Alex Szalay.

Simultaneously, the astronomy *literature* has been digitized by people like Michael Kurtz and Alyssa Goodman.

# Data preservation

Media wear out

Software becomes obsolete

Individuals retire

Institutional frameworks needed for preservation; most scientists do not think in terms of long-term preservation.

This is a social problem: scientists are more rewarded for collecting new data than for exploiting old data.

# New collaborative methods

Open access journals

Open source

Wikis

Virtual laboratories and observatories

“collaboratories”

What kinds of economic, legal and social arrangements will be needed?

Different sciences have different ethics, principles, and behavior.

# 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.

Traditionally, stuff in libraries and archives was created by people and reflected their intellectual effort ... neither libraries, archives, nor museums have historically saved non-creative activities (with a few apologies to herbaria and the like).

# 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).*

# Data is at risk from

Materials failure; acetate tape slowly turns to vinegar, magnetic layers delaminate, ... but this isn't the biggest problem.

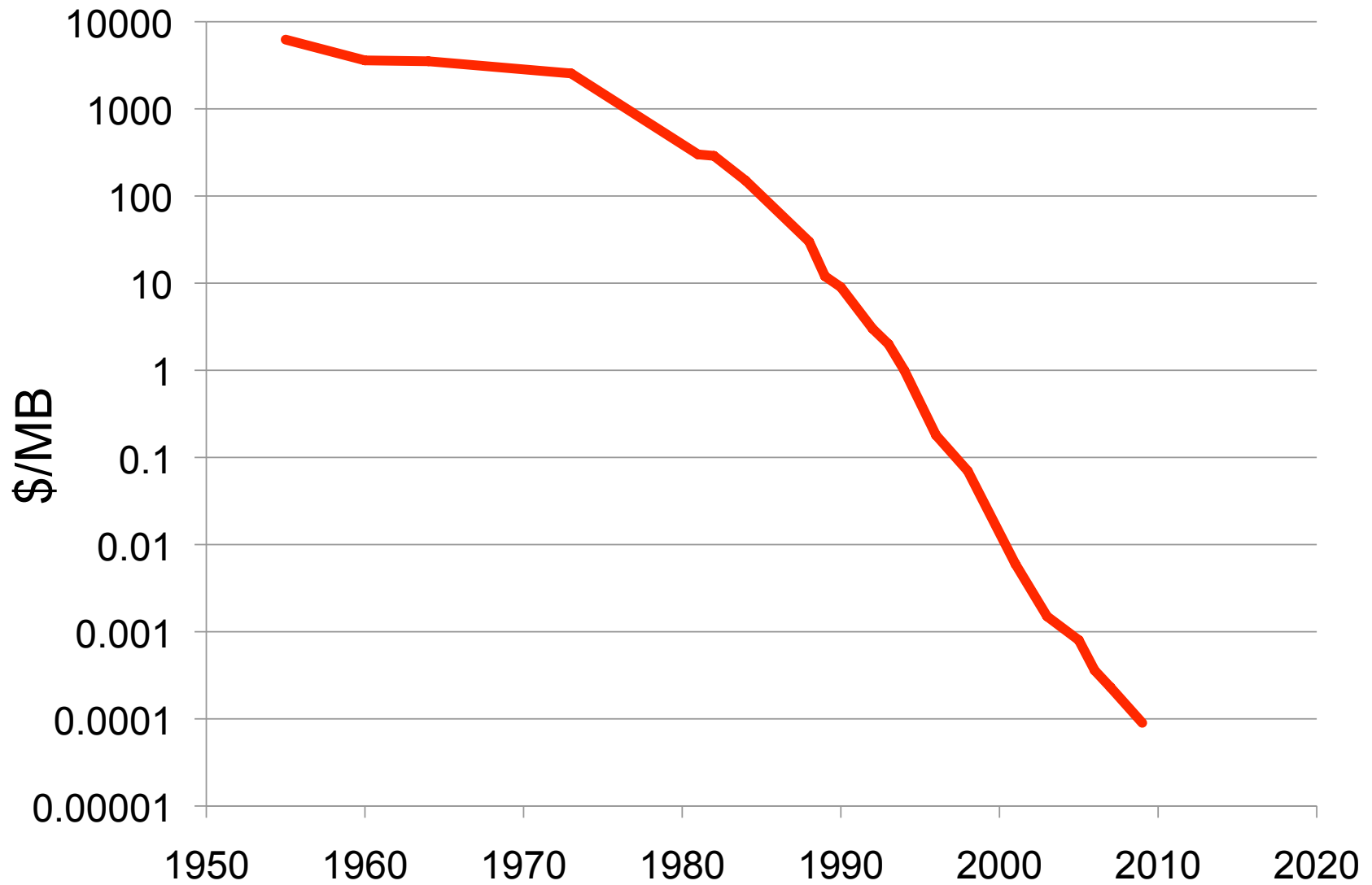
Hardware obsolescence: where would you find a 9-track tape drive today? Computers today ship without 3.5" diskette drives, despite the billions of those sold.

Software obsolescence: could you find a copy of Visicalc? Nota Bene? Could you find a machine which would run them?

Forgetting where something is.

Forgetting **what** something is. Strings of bytes aren't very useful if you don't know what they mean.

# The history of disk prices



# Historically

We published the important conclusions and the key data in journals, and libraries kept the journals.

The original data might be kept in the notebooks of a researcher, which might be saved upon retirement; everything depended on the judgment of the researcher, the other folks in the university department, and/or the archivist.

But at least if you had the notebooks, you could probably read them.

Today, a ten-year old computer medium is probably a forensic or archeological problem.

# 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.

# And there's lots of this data

Physically: “The committee estimates that more than 15,000 miles of cores and cuttings, well over a quarter of a billion line-miles of seismic data, and more than 100 million boxes of fossils are in geoscience repositories today. “ *Geoscience Data and Collections: National Resources in Peril* (National Academies Press, 2002)

Electronically: (NOAA data centers)

FY1980	1 TB	\$24.6M budget
FY1994	220TB	\$20.0M budget
FY2004	3,600TB	\$27.6M budget
FY2010	44,000TB	

# 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 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.

# 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.

# Face search (Google Picasa)

Picasa Web Albums - Add Name Tags - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Gmail Calendar Documents Photos Reader Web more

Settings | My Account | Help | Sign Out

Picasa™ Web Albums My Photos People Explore Upload Search

Name tags: All People Start from the beginning

People: All People Album: Friends

Use the checkboxes below to select faces of one person. Select: All None

People in my photos View All

Hava Anna Jaime

Mom Dad Jamie

Enter name:  Apply

Don't want to name a face?  
Mark selected faces as: Ignore Skip Not a Face

Suggestions: Jamie Mom Hava [Choose](#)

Name tag status: You have tagged 171 of 264 faces total and 89 of 147 in this album

[Edit hidden faces \(53\)](#) (faces marked ignore, not a face, or skipped)

Done

1934



1985



1965



1991



2009



This industrial portion of east Bridgeport became almost vacant land and is now starting to see single-family homes again

# Bridgeport

# Why isn't it kept?

“A large amount of valuable scientific data gathered with federal funds is never archived or made accessible to anyone other than the original investigators, many of whom are not government employees. In many instances, the organizations and individuals that receive government contracts or grants for scientific investigation are under no obligation to retain the data collected, or to place them in an accessible archive at the conclusion of the project. Thus, data sets that commonly are gathered at great expense and effort are not broadly available and ultimately may be lost” - *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (National Academies Press, 1995)

# 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)

# What do museum curators do?

They collect. The Smithsonian (National Museum of Natural History) already has 126 million items. (below is from 2007)

Curator of Anthropology: J Daniel Rogers; collects in Mexico; the collection already has 2 million items.

Curator of Botany: Warren Wagner; collects on Pacific islands; collection already has 4.5 million items

Curator of Entomology: Ted Schultz; collects in the Amazon; collection already has 32 million specimens

Curator of Invertebrate Zoology: Rafael Lemaitre; collects in the Caribbean; already has 35 million items.

Curator of Minerals: Glenn MacPherson; collected in Antarctica; already has 17,000 meteorites.

Curator of Paleobiology: Scott Wing; collects in western North America, Pakistan, Argentina, already has 40-50 million fossil plants.

Curator of Vertebrate Zoology: George Zug, collects in New Guinea, already has 500,000 amphibians and reptiles.

# Same story in New York

The American Museum of Natural History already has 32 million items. Their lead curators in 2007:

Paleontology: Mark Norell, visits Northern China.

Anthropology: Charles Spencer, visits Oaxaca

Invertebrate Zoology: Randall Schuh, collects in Australia

Vertebrate Zoology: Nancy Simmons: Amazonian Peru

Physical Sciences: James Webster, (works locally!)

# Collecting is valuable

I'm not saying we shouldn't collect. If you read through the biographies of all the curators on the last two slides, you'll find good reasons why they are gathering their various materials.

But as long as the high-status activity in Egyptology is transferring material from a hole in the ground in Egypt to a hole in the ground in London, ignoring untranslated papyri around the globe, nobody is really motivated to solve the problem.

# 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)..

# Keeping is cheaper than re-creating

“At the 1995 meeting of the ISO Archiving Standards working group it was reported that it cost (including labour) about \$5-7 per megabyte per year to retain electronic records created in the engineering sector, but about \$1250 to reconstruct them if they were lost or destroyed. Petroleum survey records are even more expensive to recreate. The National Archives of Australia hold 600,000 computer tapes containing oil survey data. These data are regularly re-used by oil exploration companies; recreating the off-shore data would have cost in the early 1990s AUS\$10,000 per metre or AUS\$10 billion in total..” (Seamus Ross, *Changing Trains at Wigan, NPO, 2000*).

# Social stresses are coming

Digitization gets steadily cheaper, and more and more of it is done. Searching techniques will then mean that it is easier to find results in old data, and more and more such results will start to come forward. But this is likely to be viewed as second-rate effort by the people who hire & fire; and it will be a problem when it is more productive than new research.

Will scientific culture adapt to the increased prestige of the data miners?

# 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.

# Even small costs seem high

Every museum is familiar with the problem of someone who wishes to donate material but will not or can not provide financial support for its care. Similarly, our library offers to preserve data or publish e-journals for faculty, but (for example) a charge of \$1,000 per year to run an e-journal seems high to prospective users.

OCLC runs a data storage operation: for up to 100GB the charge is \$60/GB/year, decreasing to \$15/GB/year for over 1,000 GB. This includes a good deal of redundancy and management; by contrast “box.net” charges \$100/yr for 15 GB, or about \$6/GB.

But, as OCLC notes, people compare this with current prices of \$300 per GB for a disk drive and complain.

# What organization?

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)

# Should research pay for storage?

The University of California Berkeley libraries will spend (total) \$47M this year, while research at Berkeley was \$585M in 2004. So it seems attractive to observe that a small part of the research budget would pay for the long-term storage of the research output.

Researchers are not likely to approve of that idea, thinking instead that the future should pay for its needs. But it seems unlikely that there will be a lucrative business in old data, any more than there is for out-of-print books.

The “open access” journals that charge authors (PLoS, Biomed Central) are an example – but we need the same for data.

# What format to store?

For long term storage, you'd like to convert to standard formats; and you'd like to do this while the people who collected the data still have the processes they used fresh in their minds. But that's exactly when they are busiest, and exactly when they may still want to use whatever format was best for the specific research project, rather than convert to something more general but perhaps less efficient for the immediate task.

We need to understand how to preserve the semantics of the original data without necessarily preserving a proprietary or obsolete data format.

# Selection

The traditional archive judges what to keep and what to discard; historically, the PRO tried to keep about 10% of what they were offered. Times are changing; even a microfilming project of recent years often spent as much choosing what to film as doing the filming. Digital storage is now really cheap compared to selection. Bill Arms observed that we were likely to start dividing digital data into three piles: (a) material of such great value that we would somehow find the money needed to preserve it, (b) material judged sufficiently worthless that we could discard it , and (c) a middle pile of stuff that we would put in a freezer and hope that our successors had better automated tools to deal with it. He suggested that the middle pile might be 90% of the whole!

# What to do?

We need

technical progress

economic progress

social progress

# Technical needs

Public data standards, perhaps area-specific semantics for XML.

Formal query language standards so that we could catalog and describe “dark web” sites

Expanding efforts like LOCKSS that save the bits (by sharing them among multiple archives)

# Research needs

Provenance tracking systems: many tasks, such as calibration, quality control, and the like should only be done once, despite a proliferation of annotated and re-used data sets.

Self-describing data: “documentation” is less and less a part of modern software and datasets.

Interoperability: different subject areas may want to share the same data.

Data in new media: visual data, visualization of data, and similar issues for sound, moving images, vector graphics, and other formats.

# 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.

# LOCKSS



LOCKSS, a project of Vicky Reich (Stanford) and David Rosenthal (Sun) helps preserve digital files by sharing them around libraries.

What is remarkable about this project is that it relies on delay – making some actions deliberately slow, to give humans time to respond and to frustrate vandals. For example, finding one copy of a work happens quickly; to find all copies will take weeks. This is why their symbol is a turtle.

It is also a cooperative with no central control – a voting scheme manages everything.

# Economic needs

Support for organizations that save and preserve data; this must survive the existence of the research funding that created it.

Formal agreement on digital rights management languages

Either economic or cultural agreement on what digital rights management in data should look like; I would very much like to see a bargain that full public access in a reasonable time is essential in exchange for any protection.

# Volunteers

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

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. One of my former students, Judith Gelernter, compared professional LCSH headings with LibraryThing tags and found the professional data better.

(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.

# Ethics/economics of data access

Some scientific areas – for example molecular biology – have rules that you can't, for example, publish the claim that you measured a protein structure without depositing the data in the public Protein Data Bank.

Other areas – such as astronomy – give the person who captured the data exclusive use for 1 or 2 years.

And yet molecular biology is of potentially immense commercial value while astronomy is not.

In many areas no tradition yet exists, and we need to work for one.

# Social needs

We need to reward the activity of collecting and mining old data.

We need to view data preservation as a responsibility of data collection, just as we view publication and correct citation as an essential part of science.

We need to train scientists in aspects of how to store data so that other people can use it.

We need a conference and journal that take papers in data mining and preservation.

# Data economics

A frequent suggestion: government should be an information wholesaler and private industry should be the retailer.

There's a lot of politics here: one book entitled *Space Economics* suggests that vendors of satellite photograph should be allowed to set their prices to try to capture the value that purchasers are gaining from it.

Issues of whether to distribute free or not have affected the SEC (Edgar), patents, and the like. The Clinton administration put through a policy to make information available at the lowest possible price (typically dissemination cost).

# 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).

Finally, in 2008 the historical Landsat archive was made available free.

# Pretty pictures



Mt. St. Helens, left 1973, right 1983. (The eruption was May 18, 1980).

# 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

<b>(billion €)</b>	<b>EU</b>	<b>US</b>
Investment value of public info	9.5	19
Economic value	68	750

# 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.

# Data quality

Can we have open and easy distribution with adequate quality? Or will we be looking for ways to fund quality as well as preservation and access? 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 – perhaps like the Digital Curation Centre's role in the UK, and 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.