



Data Management



Data deluge

Data is collected from sensors, sensor networks, remote sensing, observations, and more - this calls for increased attention to data management and stewardship



Photo courtesy of
<http://www.futurlec.com>

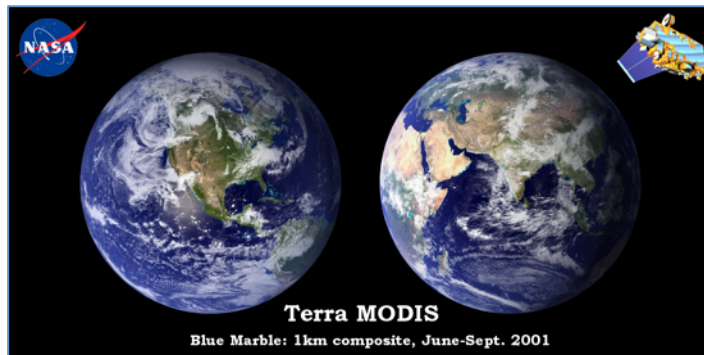


Photo courtesy of <http://modis.gsfc.nasa.gov/>



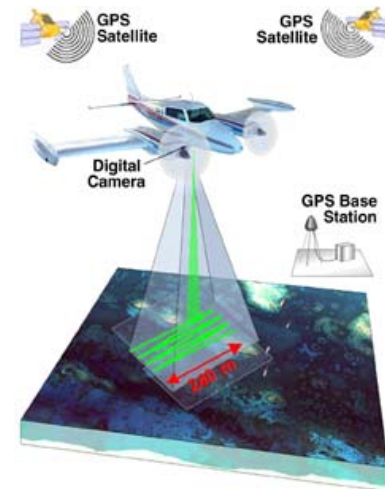
CC image by CIMMYT on Flickr



Image collected by Viv Hutchinson



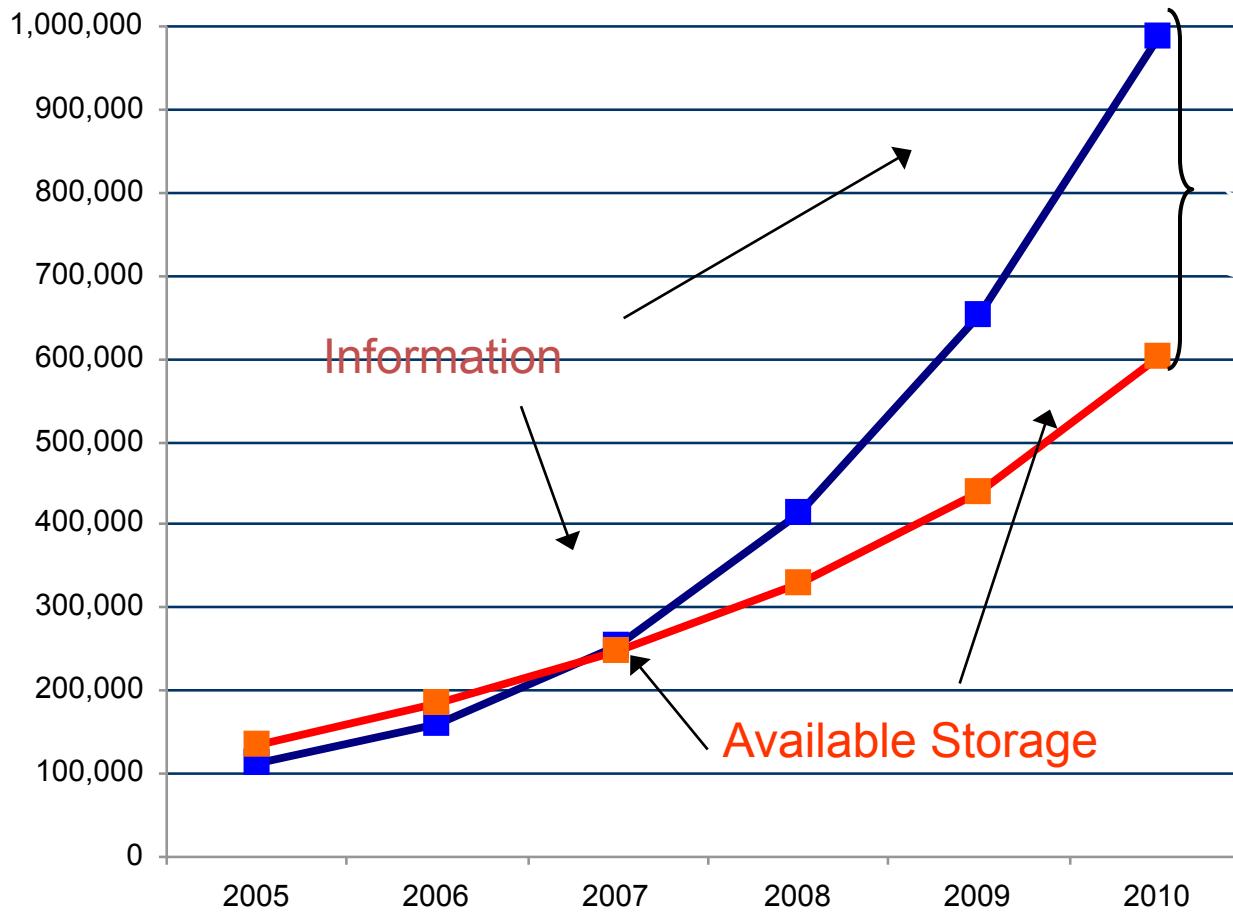
Photo courtesy of www.carboafrika.net



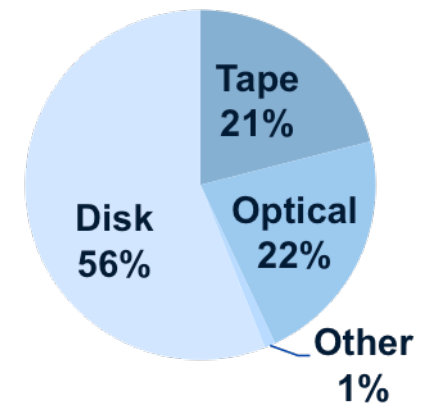
CC image by tajai on Flickr



The world of data around us



**Available Storage, 2007
264EB**



Source: John Gantz, IDC Corporation: The Expanding Digital Universe



Data loss



CC image by Sharyn Morrow on Flickr



CC image by momboleum on Flickr

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements



Example: Poor data management

A wildlife biologist for a small field office was the in-house GIS expert and provided support for all the staff's GIS needs. However, the data was stored on her own workstation. When the biologist relocated to another office, no one understood how the data was stored or managed.

Solution: A state office GIS specialist retrieved the workstation and sifted through files trying to salvage relevant data.

Cost: 1 work month (\$4,000) plus the value of data that was not recovered



Poor data management impacts everyone

“MEDICARE PAYMENT ERRORS NEAR \$20B” (CNN) December 2004

Miscoding and Billing Errors from Doctors and Hospitals totaled \$20,000,000,000 in FY2003 (9.3% error rate). The error rate measured claims that were paid despite being medically unnecessary, inadequately documented or improperly coded. In some instances, Medicare asked health care providers for medical records to back up their claims and got no response. The survey did not document instances of alleged fraud. This error rate actually was an improvement over the previous fiscal year (9.8% error rate).

“AUDIT: JUSTICE STATS ON ANTI-TERROR CASES FLAWED” (AP) February 2007

The Justice Department Inspector General found only two sets of data out of 26 concerning terrorism attacks were accurate. The Justice Department uses these statistics to argue for their budget. The Inspector General said the data “appear to be the result of decentralized and haphazard methods of collections ... and do not appear to be intentional.”

“OOPS! TECH ERROR WIPES OUT Alaska Info” (AP) March 2007

A technician managed to delete the data and backup for the \$38 billion Alaska oil revenue fund – money received by residents of the State. Correcting the errors cost the State an additional \$220,700 (which of course was taken off the receipts to Alaska residents.)

Importance of data management

guardian.co.uk

Search Environment Search

News Sport Comment Culture Business Money Life & style Travel Environment TV Video Community Blogs Jobs


Environment Hacked climate science emails

Climategate scientists cleared of manipulating data on global warming

Muir Russell report says scientists did not fudge data, but they should have been more open about their work

- Read the full text of the review here
- 'Climategate' report - main findings

David Adam, environment correspondent
The Guardian, Thursday 8 July 2010
Article history



Muir Russell during the release of his report into the scandal of the hacked emails sent by climate scientists from University of East Anglia. Photograph: Sang Tan/AP

The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics, but a review found they had failed to be open about their work.


Sir Muir Russell, the senior civil servant who led a six-month inquiry into the affair, said the "rigour and honesty" of the scientists at the Climatic Research Unit (CRU) at the University of East Anglia (UEA) were not in doubt. His investigation concluded they did not subvert the peer review process to censor criticism and that key data was freely available and could be used by any "competent" researcher.

Envir
Hacked
emails
Clima
scepti
Scien
Clima
Educ
Unive
Highe
UK n
Tech
Hacking

More news

More on this story

UEA's delayed response to climate emails caused by shock, says professor
Former head of research unit responds to criticism by arguing for necessity of assessing excerpts by



The Economist


SPECIAL

ENGAGE YOUR WORLD. Start Now with 12 issues for just \$12

On Environment

Most viewed Zeitgeist Latest

Last 24 hours



1. BP oil spill mostly cleaned up, says US

2. Battle to halt BP oil spill is nearing its end, says Barack Obama

The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics, but a review found they had **failed to be open enough about their work.**



Why manage data: Value to self

- Stay organized
 - be able to find your files (data inputs, analytic scripts, outputs at various stages of the analytic process, etc)
 - identify easily versions that can be periodically purged
 - Track your science processes for reproducibility
 - Quality control your data more efficiently
- Prevention of loss
- Sharing data allows you to gain credibility and recognition for your science efforts



Why data management: Advancement of science

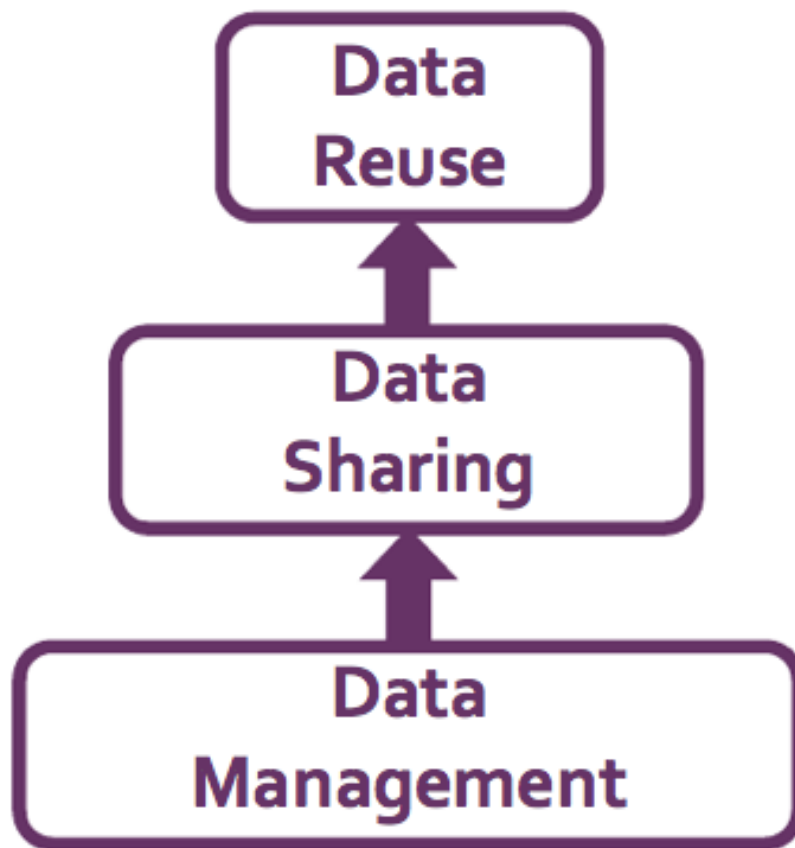
Good data management ...

- Ensures sustainability and accessibility in long term for re-use in science
- Increases the impact and visibility of research
- Promotes innovation and potential new data uses
- Leads to new collaborations between data users and creators
- Maximizes transparency and accountability
- Enables scrutiny of research findings
- Encourages improvement and validation of research methods
- Reduces cost of duplicating data collection
- Provides important resources for education and training



Benefits of good data management

Facilitates sharing and re-use...





Re-use, integration and new science

eBird



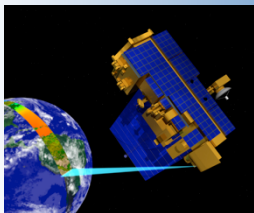
Land Cover



Meteorology



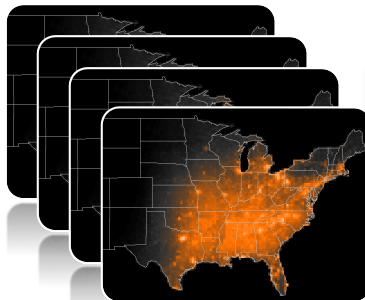
MODIS –
Remote
sensing data



Bird observations and
environmental data from >
350,000 locations in US
integrated and analyzed using
High Performance Computing
Resources



XSEDE
Extreme Science and Engineering
Discovery Environment



$$F(X, s, t) = \frac{1}{n(s, t)} \sum_{i=1}^m f_i(X, s, t) I(s, t \in \theta_i)$$

Spatio-Temporal Exploratory
Models predict the
probability of occurrence of
bird species across the United
States at a 3 km x 3 km grid.

Model results

Occurrence of Indigo Bunting (2008)

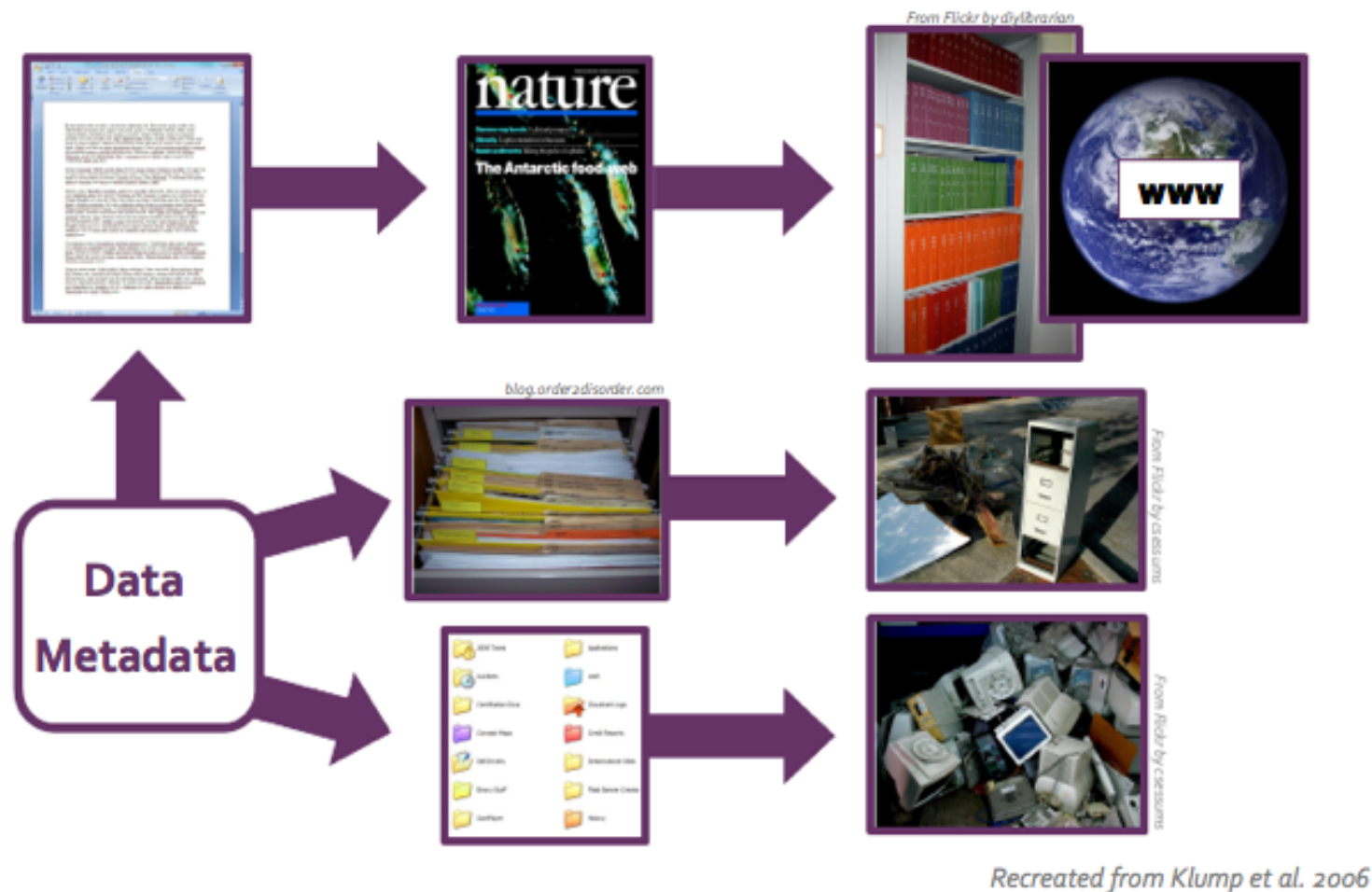


Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat use
- Measure population trends

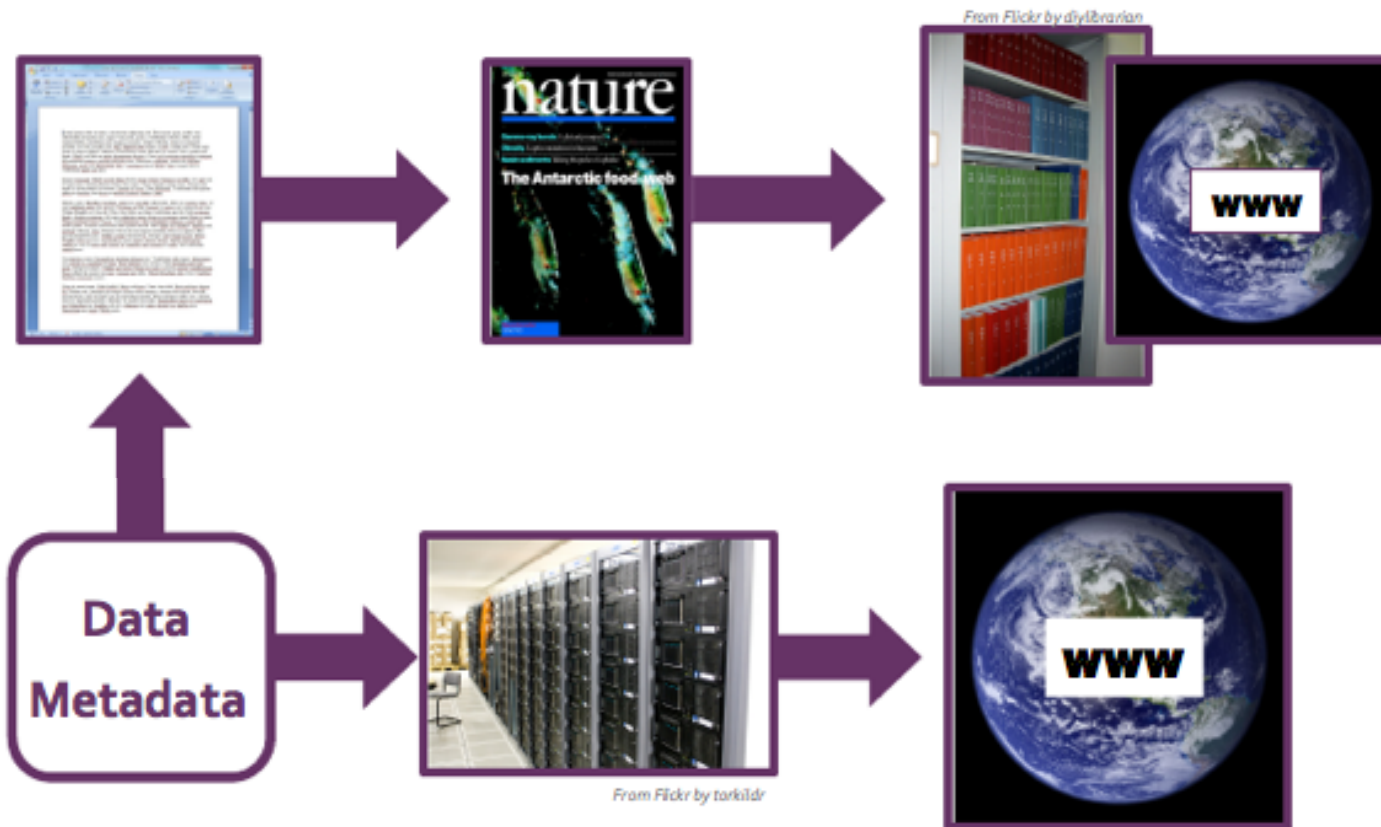


Where majority of data ends up





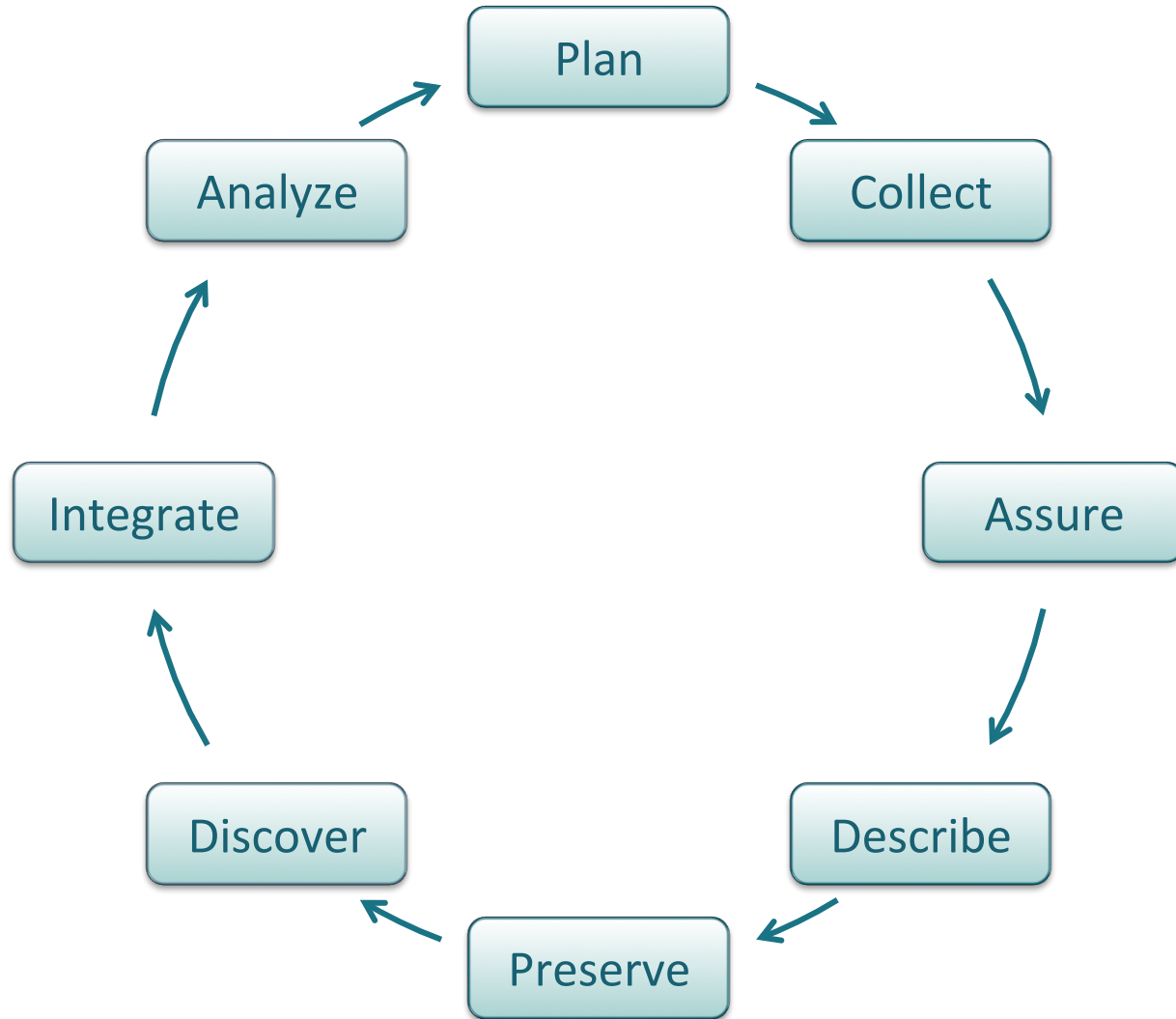
Alternative



Recreated from Klump et al. 2006



The data life cycle





Data management summary

- If data are:
 - Well-organized
 - Documented
 - Preserved
 - Accessible
 - Verified as to Accuracy and validity
- Result is:
 - High quality data
 - Easy to share and re-use in science
 - Citation and credibility to the researcher
 - Cost-savings to science

Data Sharing





Data sharing and the data life cycle

Several stages require critical attention to ensure effective data sharing

Describe

document the data content, character and process

Deposit

store the data in a location from which it can be accessed

Preserve

select storage formats and media with long term use in mind

Discover

publish information about the data so that others can find it



Value of data sharing

Public

- A better informed public yields better decision making

Sponsor

- Data sharing enhances the value of research investments

Community

- Build upon the work of others and further science

Individual

- Receive recognition for their work
- Greater opportunities for collaboration



How to make data sharable

- Create robust metadata that is discoverable
- Include archival and reference information
- Have data contributors review your metadata to ensure validity and organizational ‘correctness’?
- Publish your metadata via:
 - Data Portals / Clearinghouses
 - Federal
 - Other Online Resources



Data sharing summary

- Data sharing adds value to the data
- It is the responsibility of the researcher to share their data
- Metadata supports data accountability, liability, and usability
- Sponsors expect, some require, data to be shared
- Data sharing is essential to the advancement of science

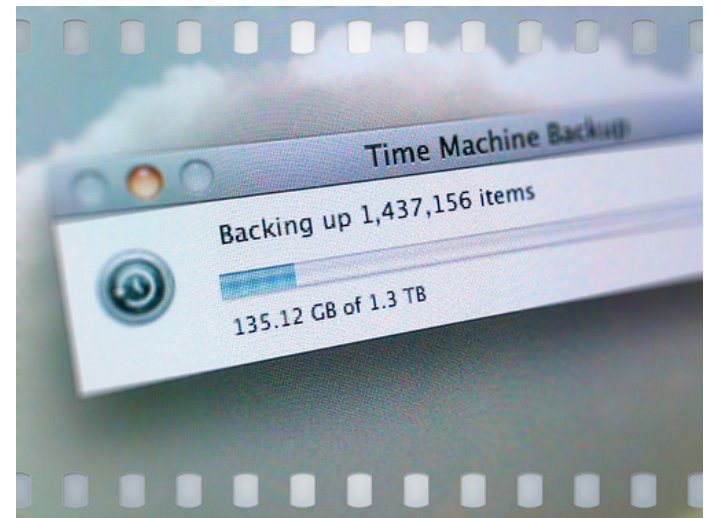


Data Preservation



Why Preserve Your Data

- Limit or negate loss of data
- Save time, money, productivity
- Help prepare for disasters
- Reproduce results of past procedures
- Respond to data requests
- Limit liability





Considerations

- How often should you do backups?
- What kind of backups should you perform?
- What about non-digital files (such as papers)?
- Where will you store your files?
 - Personal external disk
 - Centralized computer storage
 - Data repository
 - Cloud storage
- What metadata is needed when using these systems?



Other Considerations

- Data Conversions and Formats
- Versioning
- File Naming
- Create a comprehensive backup



Repository Considerations

- Are there replicas of the data?
- How long do you/they keep the data?
- What happens to the data after the project is no longer funded, project ends, or staff departs?