ETD 2011
SOUTH AFRICA

Data Curation Workshop

16 September 2011

Suzie Allard Ph.D.
University of Tennessee, USA
Edward A. Fox, Ph.D.
Virginia Tech University, USA
Lucia Lötter, Ph.D.
Human Sciences Research Council, South Africa
Lynn Woolfrey
University of Cape Town, South Africa

NDLTD
NRF National Research Foundation
CHELSA
UNIVERSITEIT STELLENBOSCH UNIVERSITY
University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

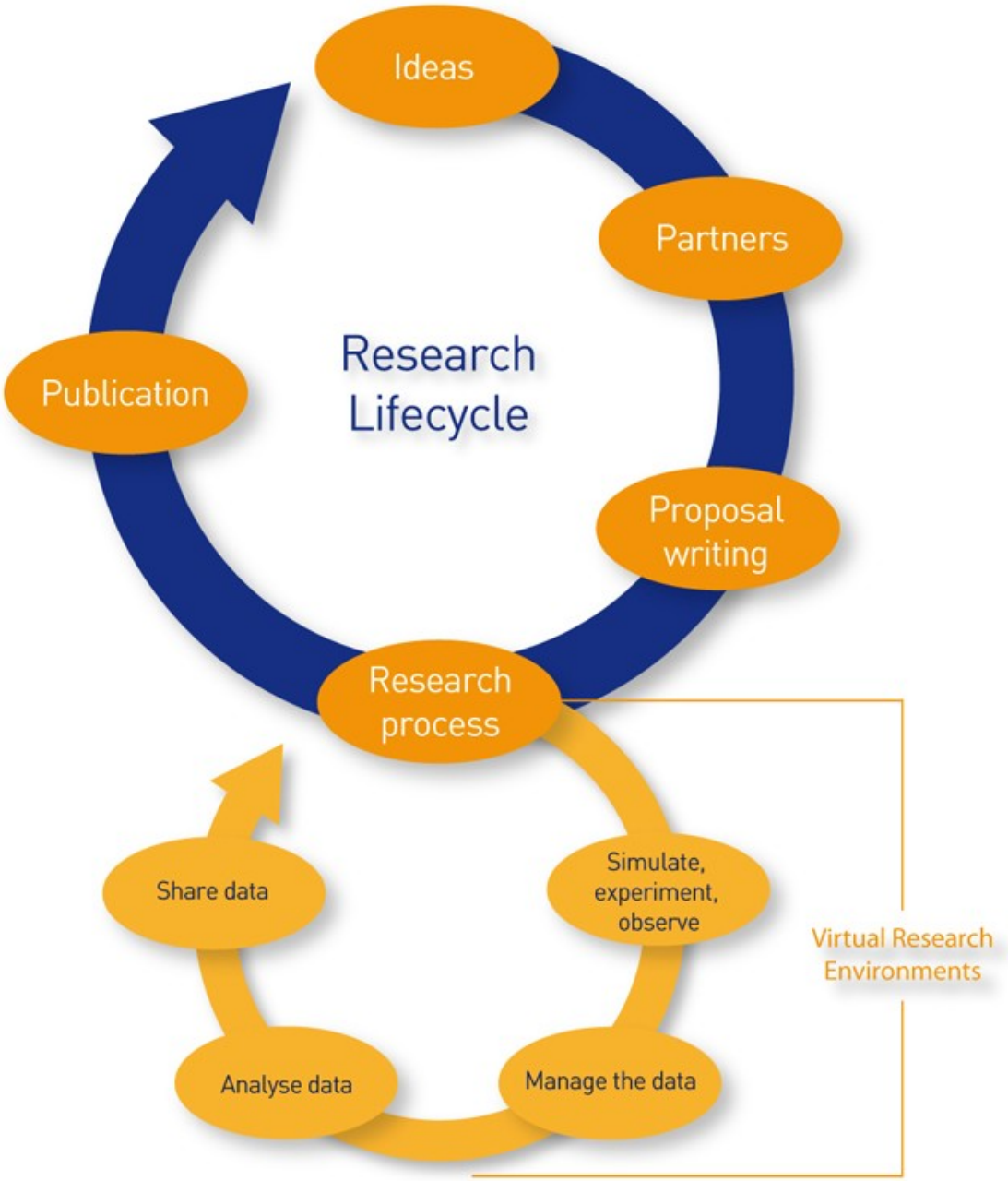
Drs. Jegede & Habib's Talks

Welcome

- Data has many definitions. But an unifying theme is that it is research output that is not codified in journal articles, monographs etc.
- Data curation is a world-wide initiative since we recognize the importance of the primary data as well as the products that come out of our data.

Introductions

The Data Lifecycle



JISC http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx

The Data Lifecycle

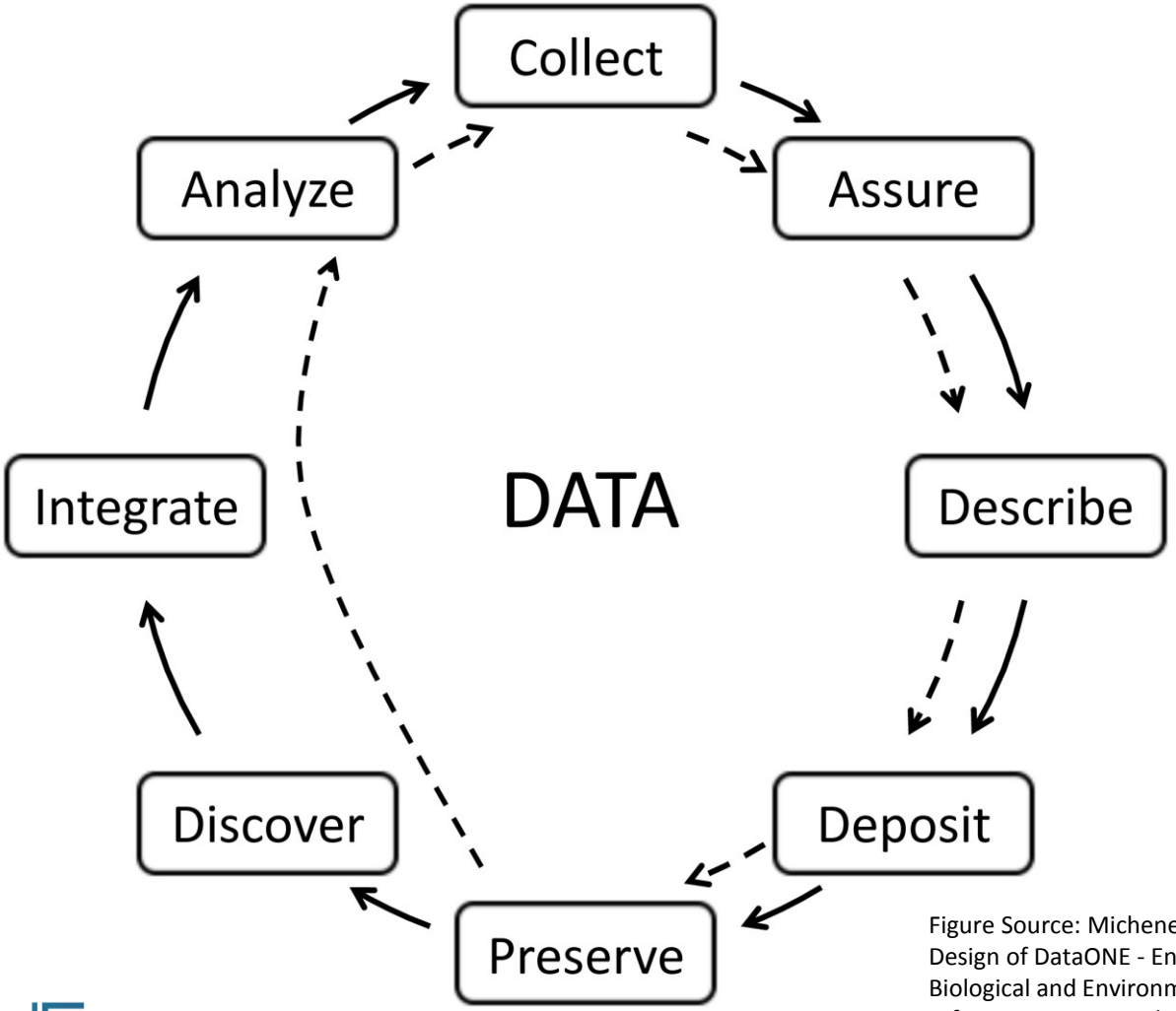
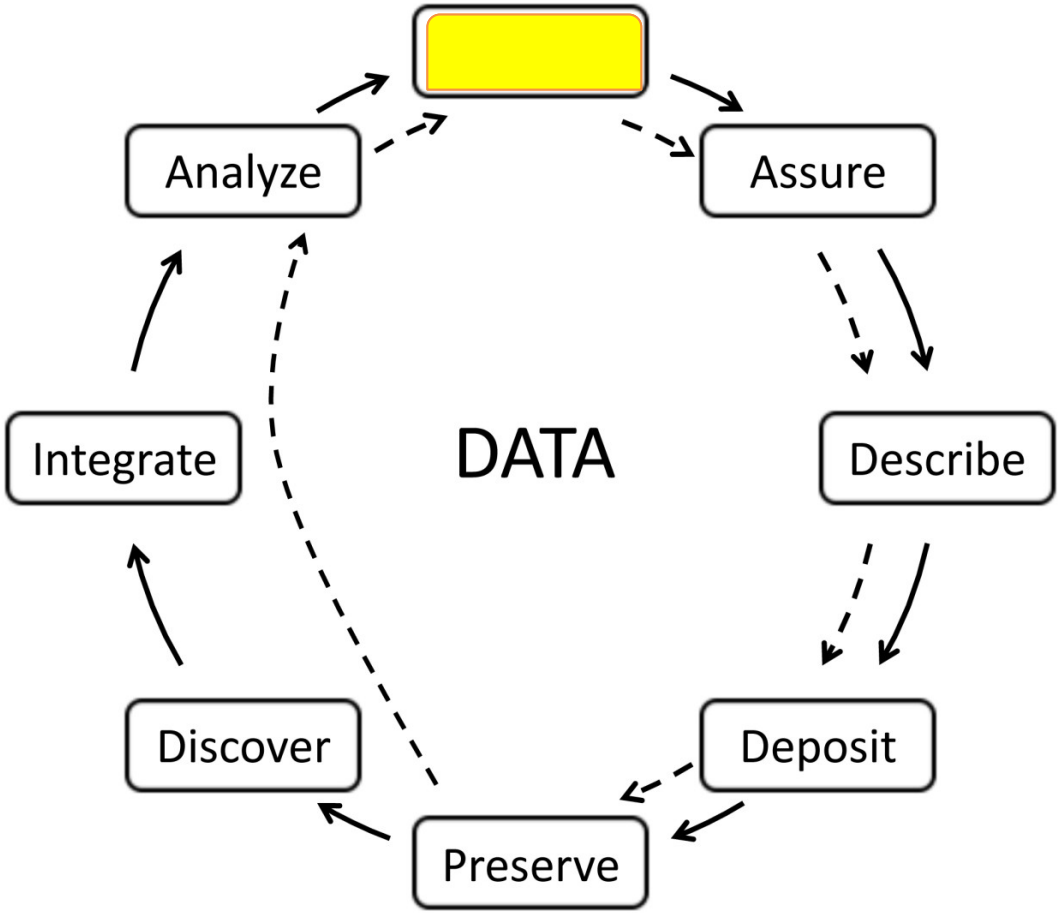
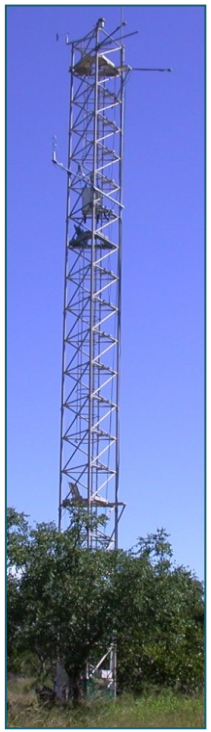


Figure Source: Michener et.al. 2011). "Participatory Design of DataONE - Enabling Cyberinfrastructure for the Biological and Environmental Sciences." *Ecological Informatics.* Accepted, in press, available online 3 September 2011. Accessed 5 September 2011. doi:10.1016/j.ecoinf.2011.08.007

DataONE

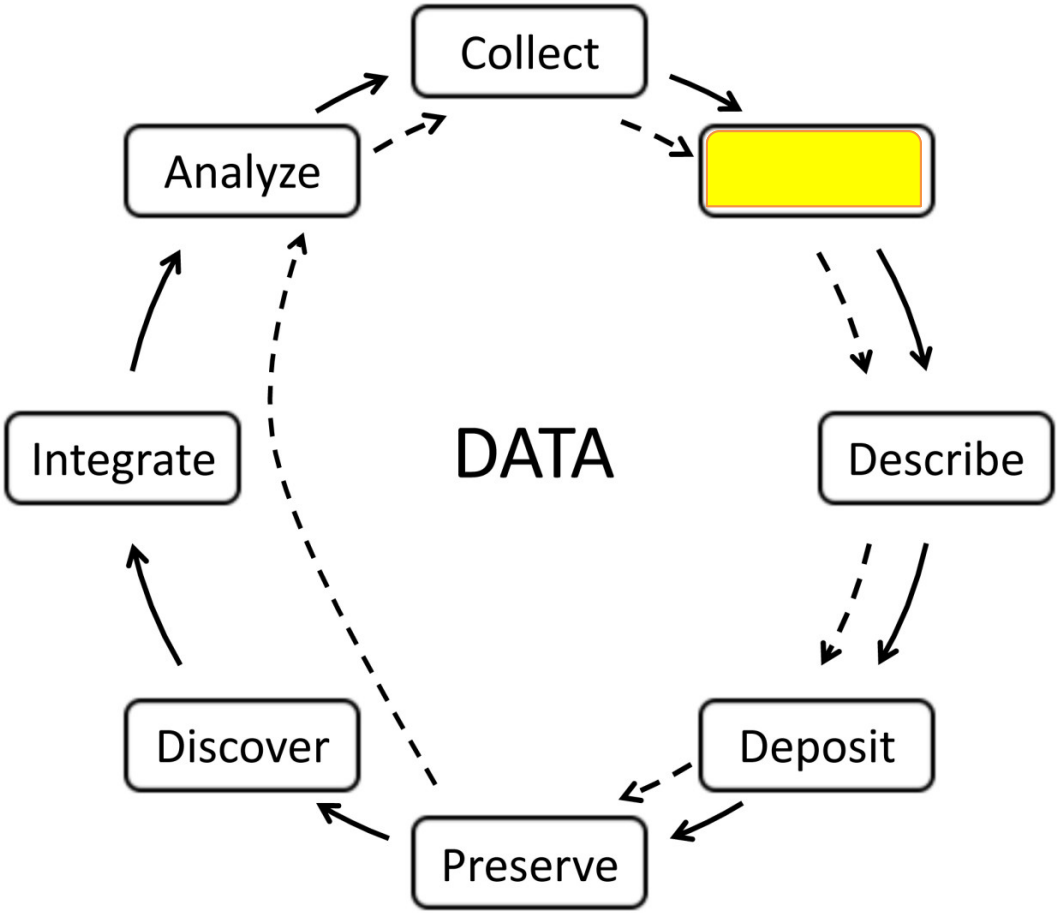


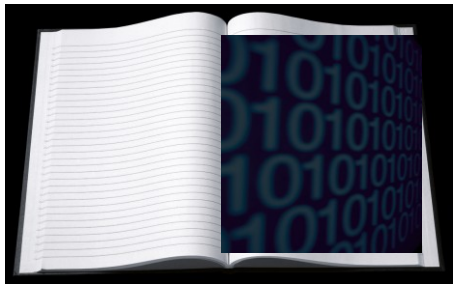


The researcher follows the research design to collect the data.

For example, in the environmental sciences this may be observations from individual researchers or from monitoring instruments.

DataONE





The researcher reviews the data to assure its quality.

This could take many forms since data may be recorded in a variety of ways ranging from notebooks to vast amounts of electronic observations.

This step is essential to providing reliable, accurate data.

DataONE

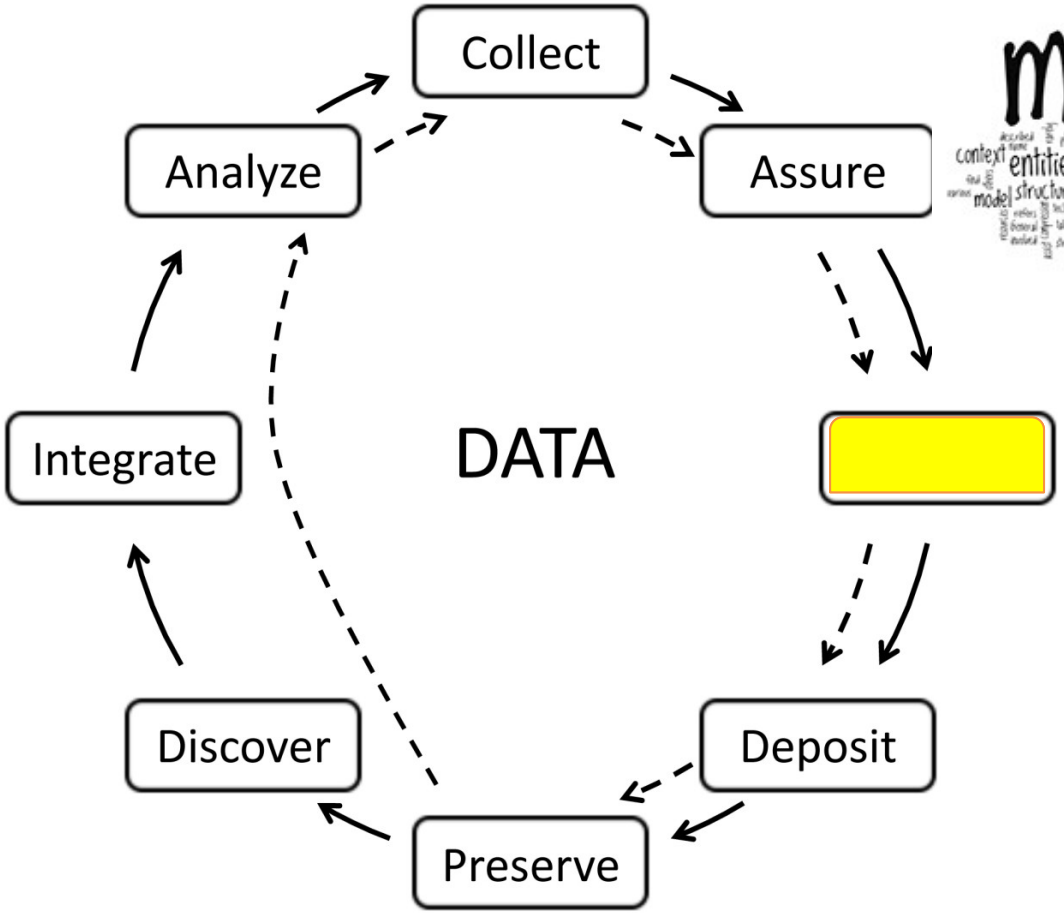


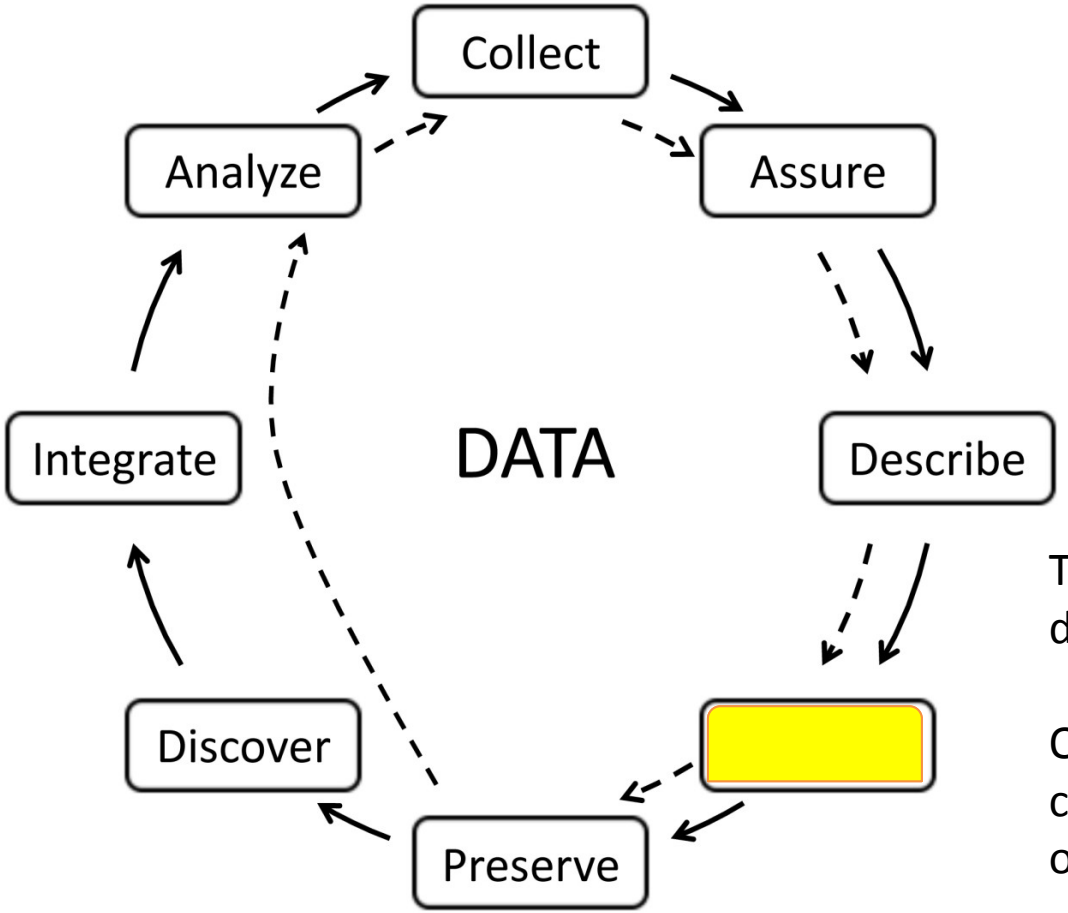
image from: http://www.biowisdom.com/tag/metadata/

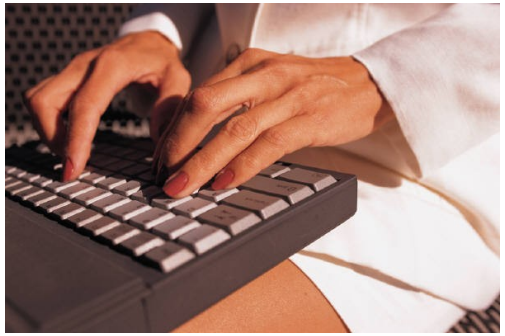
The researcher or designee describes the data using metadata.

Often researchers create their own schema rather than using standards.

This step is essential to helping discovery.

DataONE



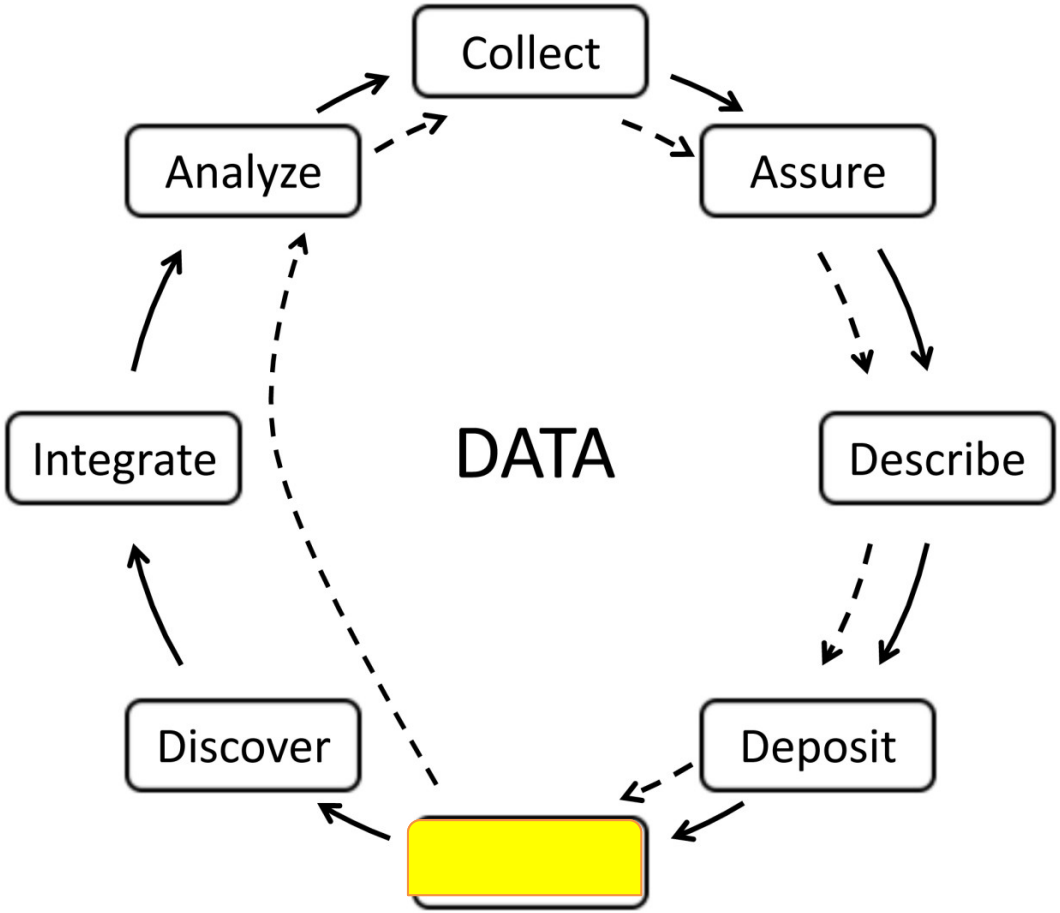


The researcher or designee deposits the data in a repository.

Often this is simply another computer in the researcher's office.

It is best if it is a curated repository that will help the researcher's data in the next two steps – preservation & discovery

DataONE



Data needs to be preserved whether it is analog or digital.

Preservation challenges include data loss, scattered data sources, data deluge, poor data practices.

DataONE

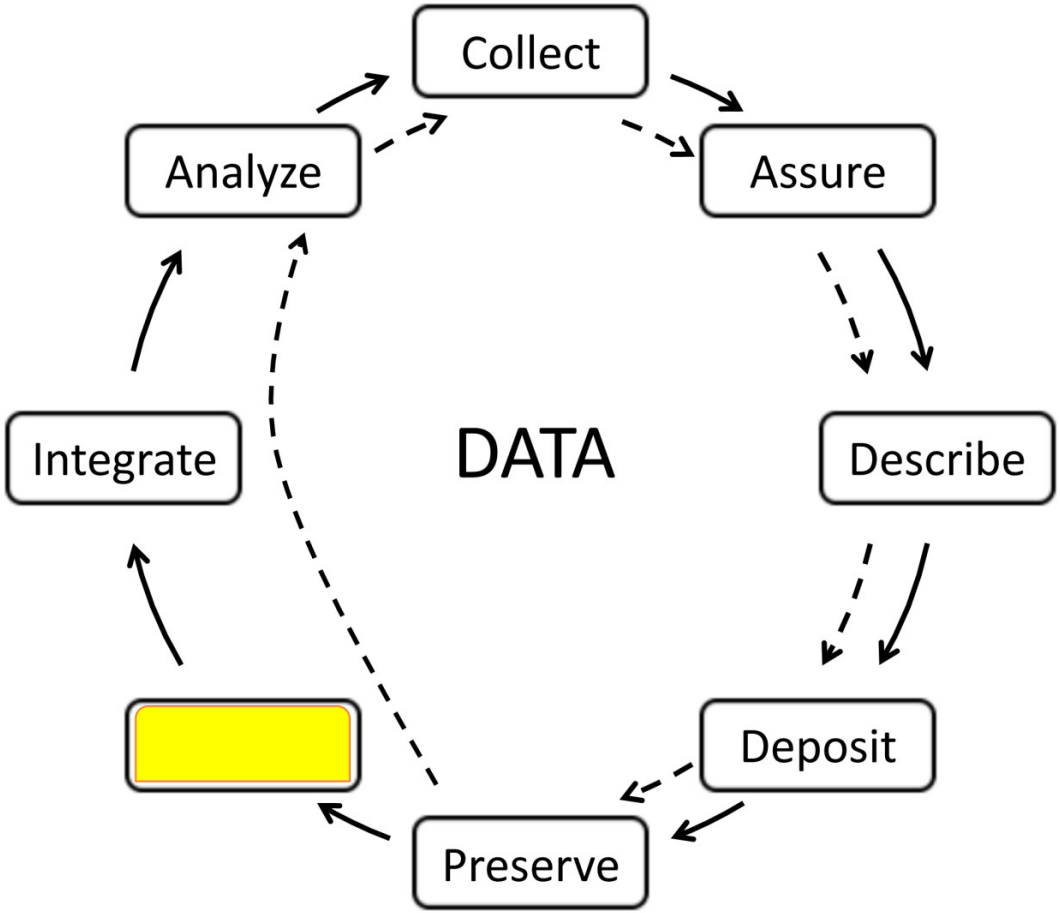


Image courtesy of DataONE

The data are “discovered” by another researcher looking for data to assist his research.

This step may also be an opportunity for data to cross disciplinary lines.

DataONE

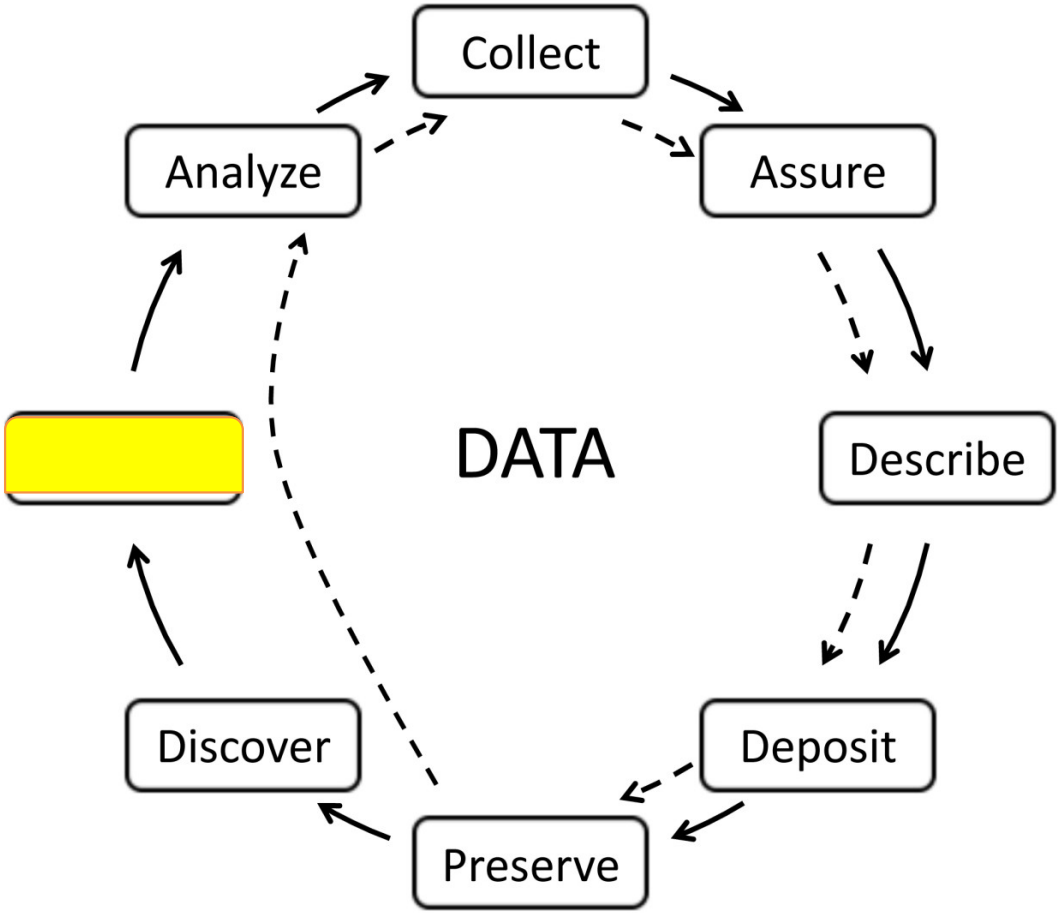


Image courtesy of DataONE

Data from different collections are discovered and integrated around one research question.

DataONE

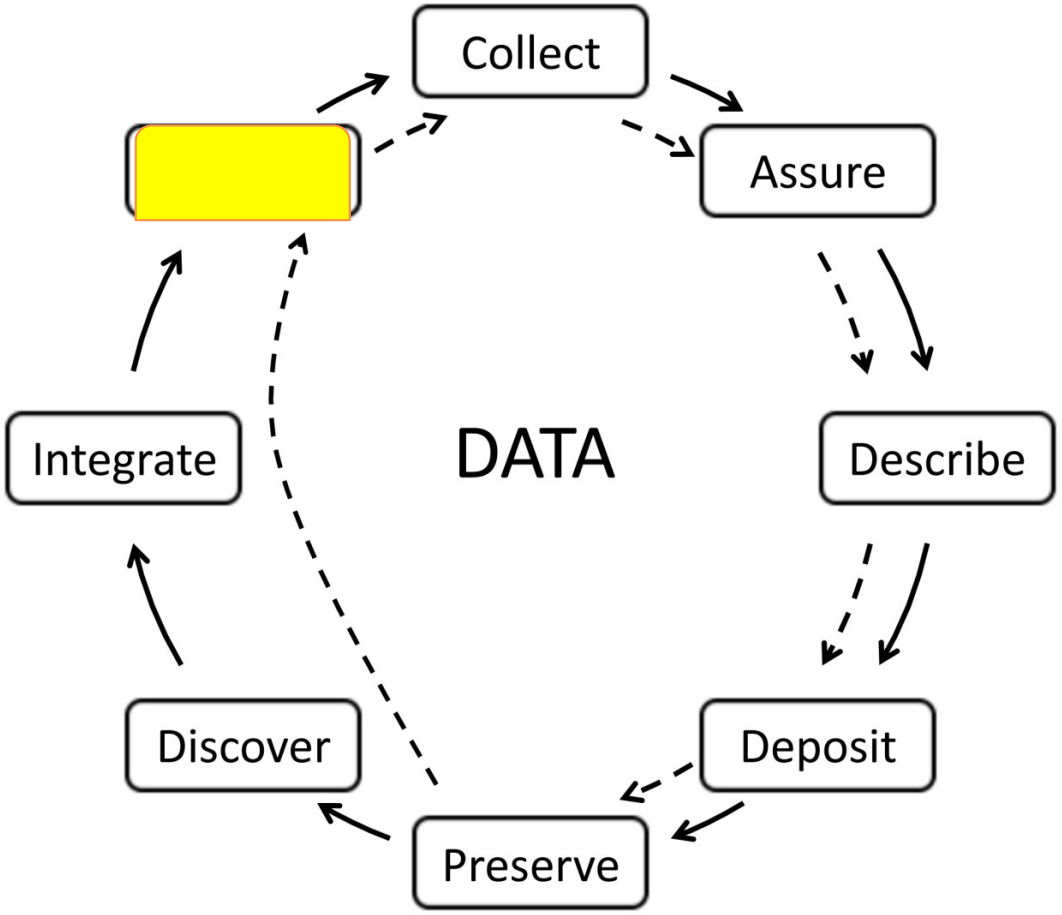
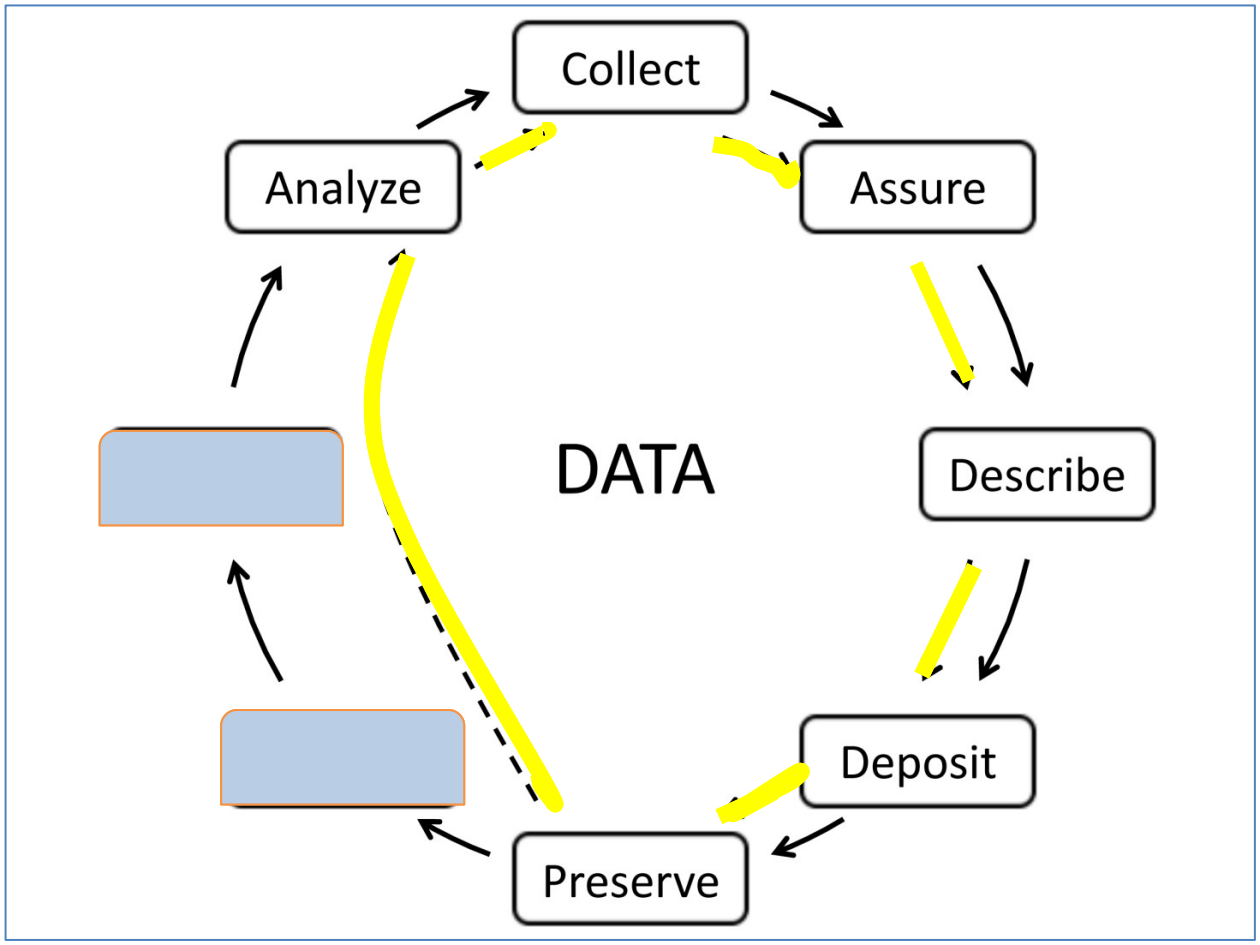


Image courtesy of DataONE

Integrated data are analyzed and new knowledge is created.

DataONE



Data may not go through all the steps in the data lifecycle.

For example data may be preserved and then be analyzed by a researcher who helped collected it.

This analysis may inform the next round of data collection.

DataONE

DataONE

DataONEpedia

Data Files Best Practices

Data & Metadata tools

Role of Data Curation

- Lucia Lötter

Data Sharing

- Lucia Lötter

Data Management & Workflows

- Lucia Lötter

Case Study: Virginia Tech

Presented by Edward A. Fox, Ph.D

fox@vt.edu .

What is Data Citation?

- bibliographic means to reference a data set
- Allows for proper attribution in a scholarly work.

Just as we cite papers, so too should we cite data sets

Four Outcomes from Data Citation

- Enables finding the data & repeating the researcher's analysis.
- Enhances visibility of the data producer & promotes scholarly recognition for their research.
- Increases visibility of the data repository → may increase use of the data holdings.
- Leads to increased data sharing supporting scientific discovery and the scientific enterprise.

Find Data, Repeat Research

- Scientific method founded on inquiry based on collecting observable and measurable evidence.
- identifying a question, doing background research, constructing a hypothesis, testing the hypothesis with data, and analyzing and reporting the results.
- Verification: Repeatable and allow for predicting future results.
- Full disclosure.

Data citation plays an important role in this verification process, as citation provides scientists with a link to the data used in the testing and analysis steps, which allows scientists to repeat the analysis, and verify the results.

Attribution

- The lack of data citation best practices is a hindrance to the scientific community, since it limits the ability for credit attribution (Altman & King, 2007).
- The need for data citation is becoming widely recognized in the scientific community.
- Even publishers are becoming active partners in encouraging a data archiving policy that includes preserving data sets and linking them to articles.

Attribution is an essential to encouraging scientists to share data. This is an important step towards following the scientific process.

Intellectual Property

- Gives credit to data producers and data publishers.
- Provides a link from the traditional literature to the data, enhancing the intellectual property link.
- Gives intellectual legitimacy to the creation of data → important scientific product worthy of being stored and cited.

Incentive to Publish, Share, Preserve

- Increases the incentive to publish a data set.
- Facilitates sharing of the data.
- Supports curation and preservation activities.

The outcome is that data is more likely to be accessed and re-used, a condition which leads to increased scholarly productivity from an initial data product.

Ideal Data Citation

- describe any data set, database, or data file.
- provide for all levels of granularity (table, row, cell).
- be used for any snapshot (version, e.g., in time)
- be formatted for any view: XML, HTML, CSV, etc.
- have (or not have) annotations
- link to older, newer, and latest versions
- provide actionability (“Click-through”)
- provide persistence (validity into the future)
- be machine readable, thus allowing for automatic parsing

Exemplar Citations

On-Line Data Set

- Turner, D.P., W.D.Ritts, and M. Gregory. 2006. BigFoot NPP Surfaces for North and South American Sites, 2002-2004. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.doi:10.3334/ORNLDAAC/750.

Subset of larger Data Set

- Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2009. MODIS subsetted land products, Collection 5. Available on-line [http://daac.ornl.gov/MODIS/modis.html] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A. Accessed November 20, 2009.

Online Map

- Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2009. FLUXNET Network Map. Available online [http://www.fluxnet.ornl.gov/fluxnet/Maps/Political_fluxnet_networks_cropped_small_april2009.png] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A.

ORNL DAAC citation examples (available at http://daac.ornl.gov/citation_policy.html

Establishing a Community of Practice

- Lynn Woolfrey

Some CoP in Education

THE UNIVERSITY of TENNESSEE KNOXVILLE

SCIENCELINKS² PH.D

SCIENCELINKS²

UNIVERSITY OF TENNESSEE COLLEGE OF COMMUNICATION AND INFORMATION

Linking Education and Science to Develop the Next Generation of Educators for Science Librarians and Data, Information and Communication Specialists

SciData

Partners:

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE The iSchool at Illinois

OAK RIDGE National Laboratory

DataONE

USGS science for a changing world

NCAR

Funded by:

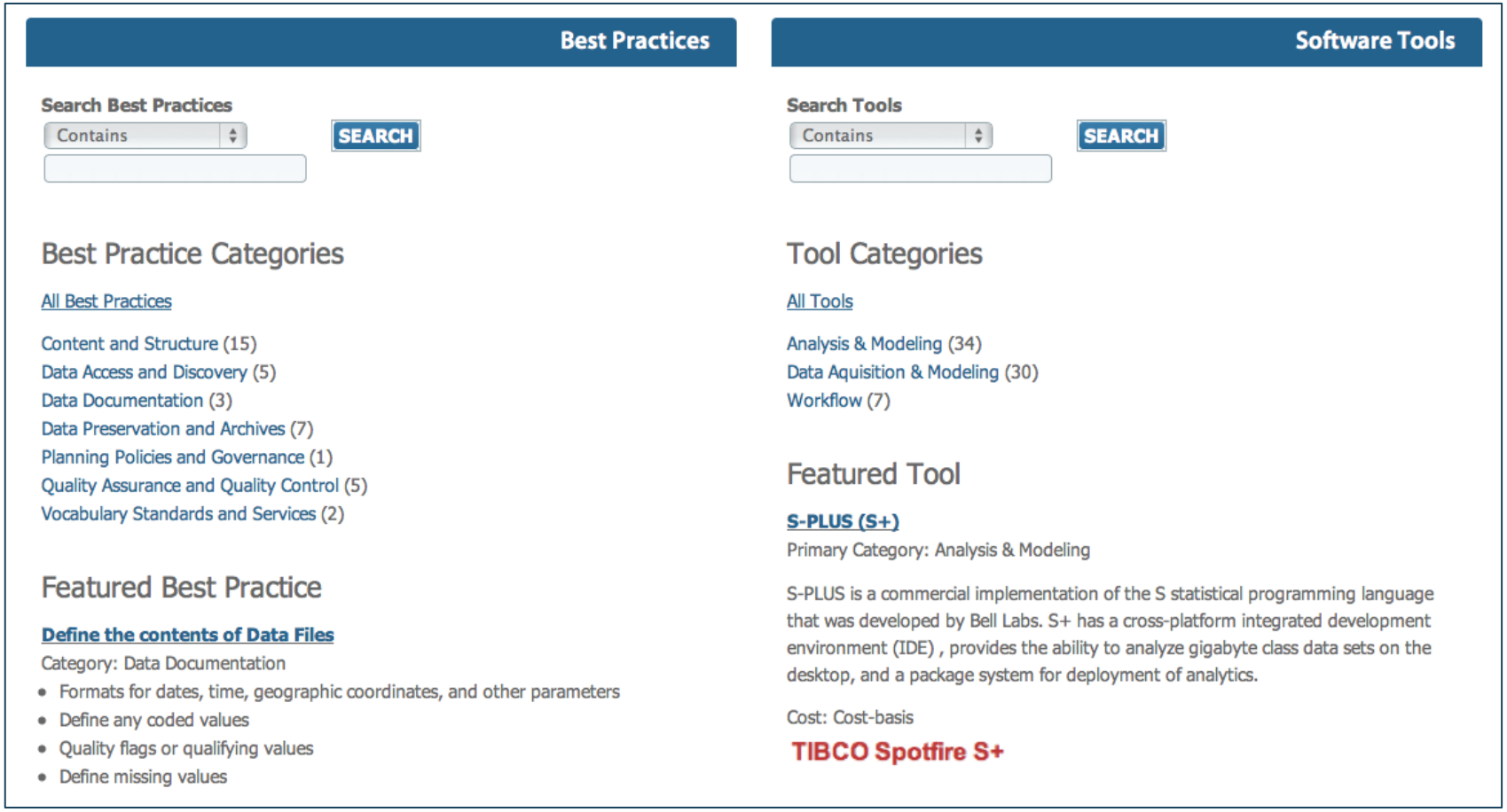
INSTITUTE of Museum and Library SERVICES

What you can do..

- Plan for how your institution will handle data
- Help make your researchers away of the importance of preserving their data
- Teach data users how to properly cite data sets they are using
- Network with colleagues to share experiences and expertise regarding data curation
- Make use of national & international resources about data citation.

Questions?

A Resource



www.dataone.org/dataonepedia
Editors: Cook, Michener
Contributors: Best practices workshop participants