Research Data Management for Today, Tomorrow, and Beyond: an Introduction to Core Principles, Practices, and Required Skills

Karl Benedict & Nancy Hoebelheinrich

kbene@unm.edu & nhoebel@kmotifs.com

Support Provided by:

UNIVERSITY LIBRARIES

Knowledge Motifs LLC
Mapping sensible data relationships

INSTITUTE of Museum and Library SERVICES

SIP
MAKING DATA MATTER

Outline

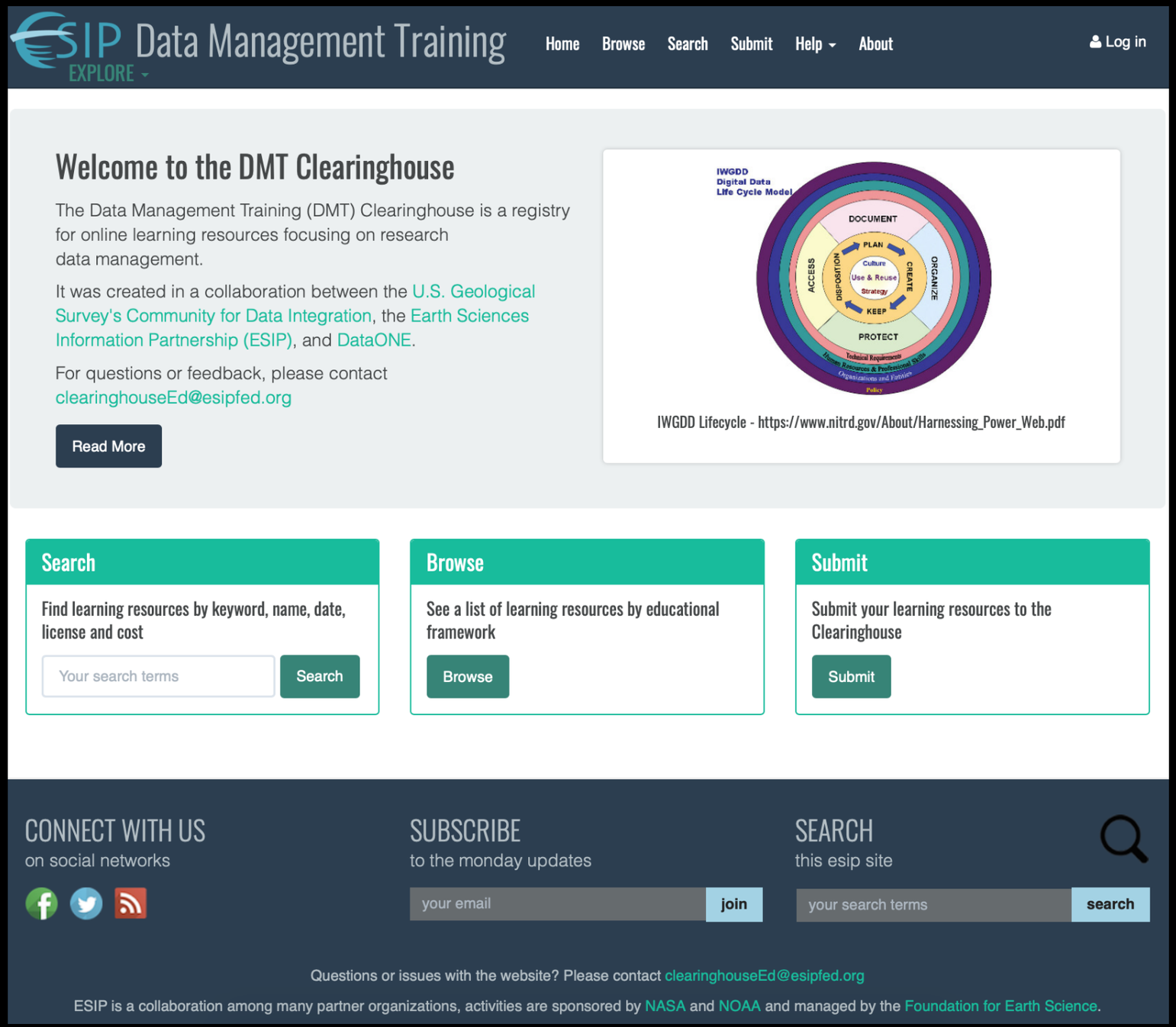
- Context
- “Data” and “Documentation”
- Some Definitions
- Some Recommendations
- A Process

Context

Data Management Drivers

- Increased research efficiency through well planned data management and collaboration
- Data Management Plans required by funding agencies
- Publications requiring links to supporting data
- Increased impact of research through **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) data

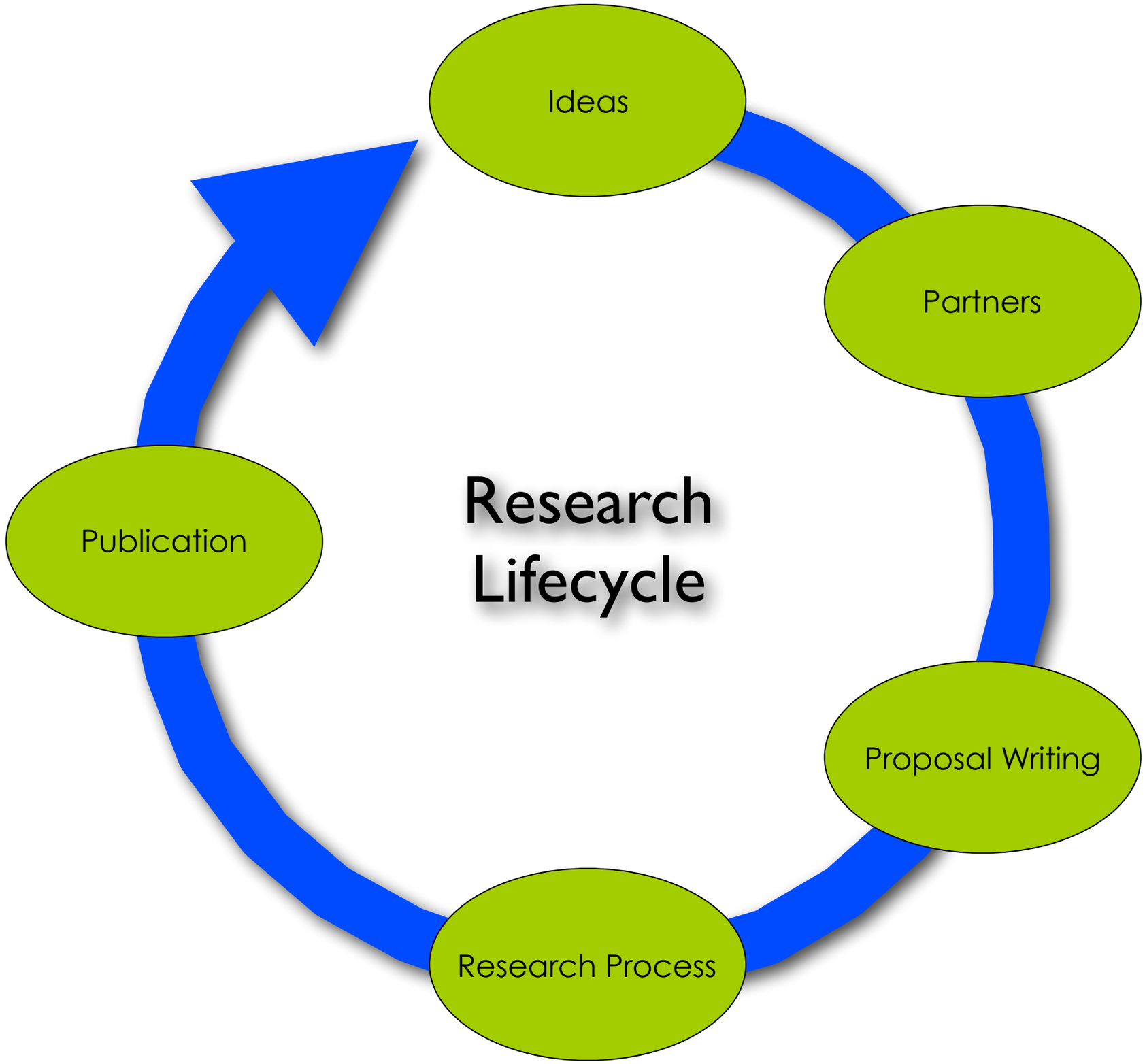
Introduction to the Data Management Training Clearinghouse



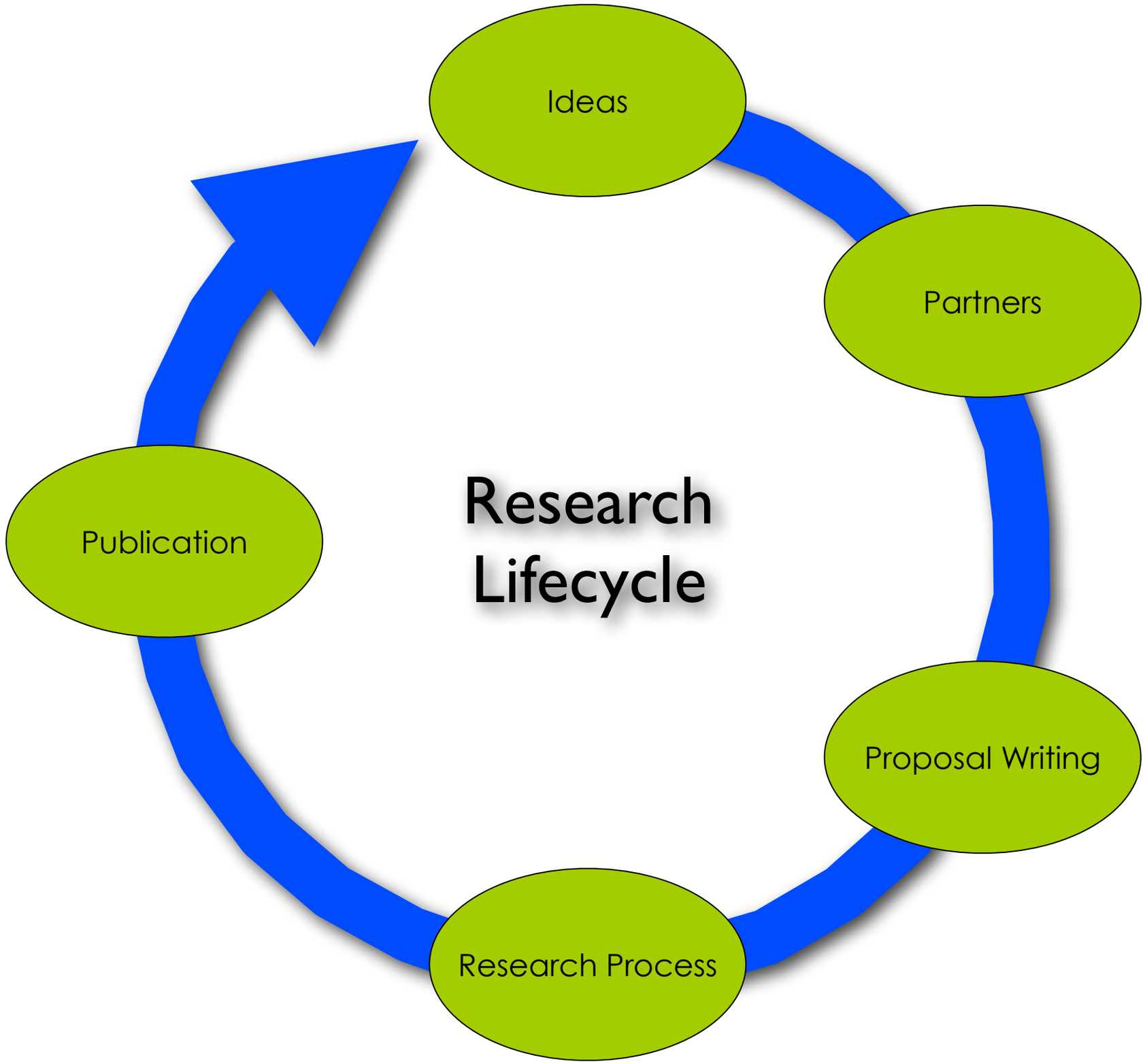
dmtclearinghouse.esipfed.org

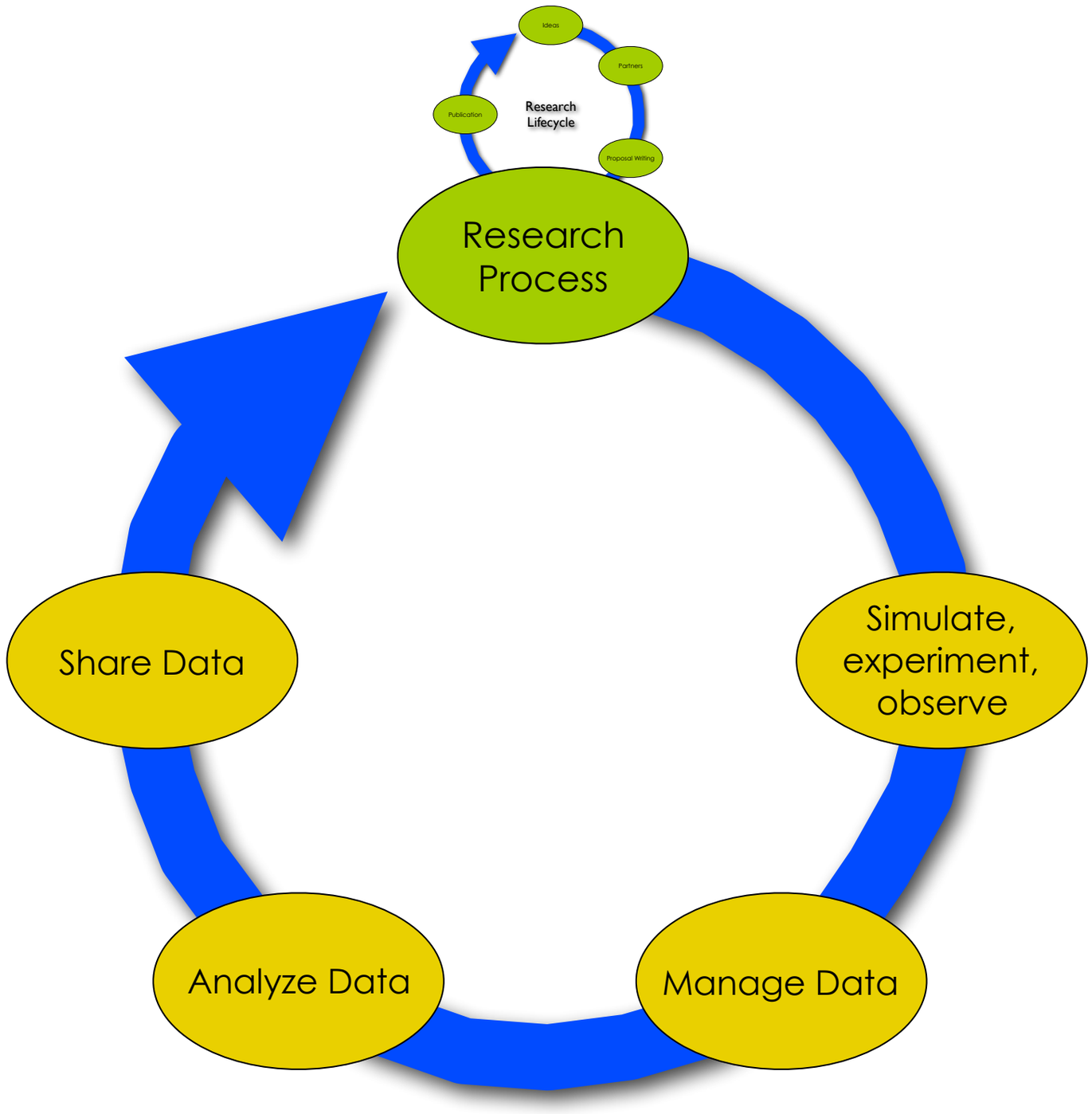
A Cross-walk between the Research and Data Lifecycles

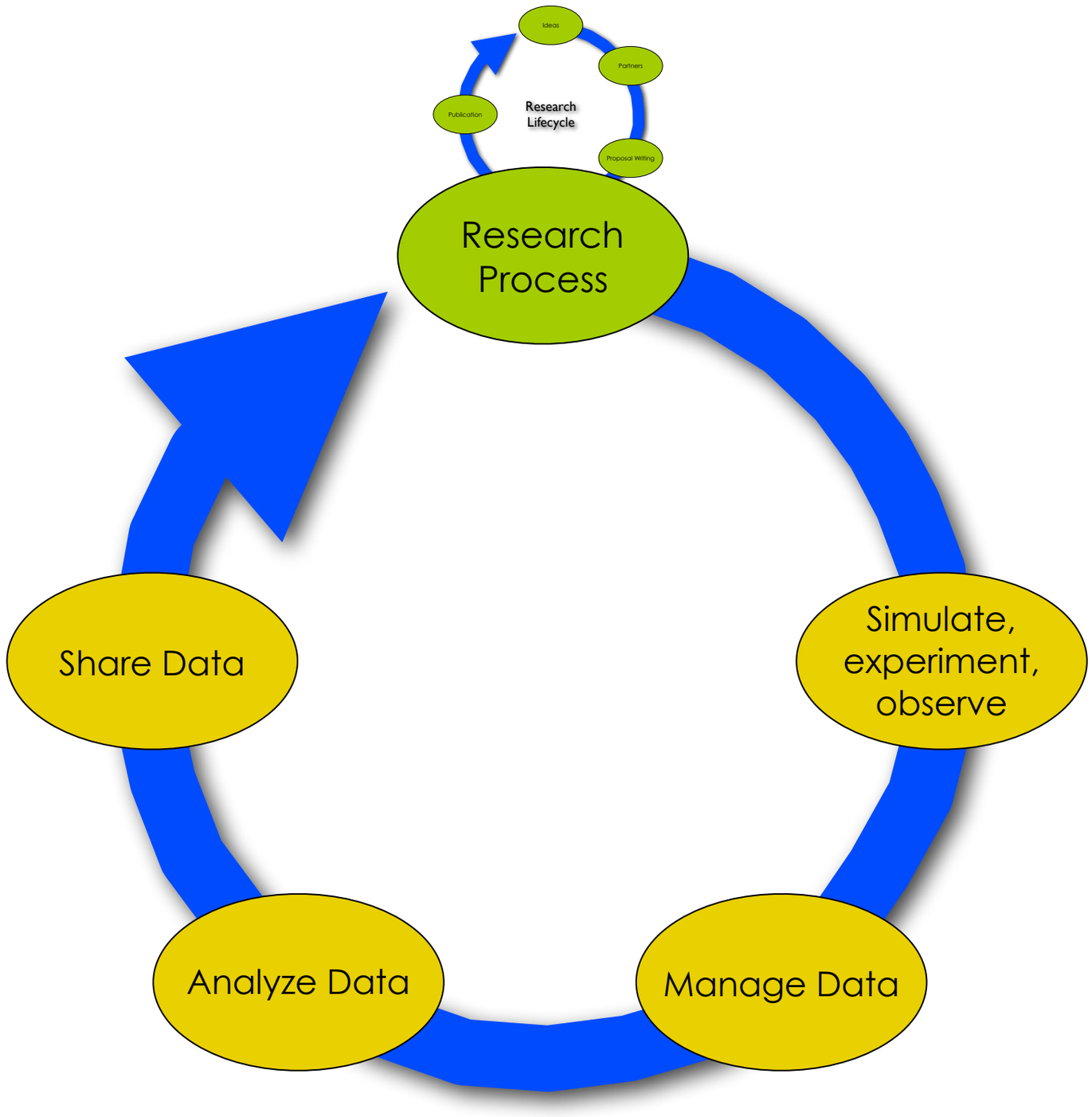
Research Lifecycle

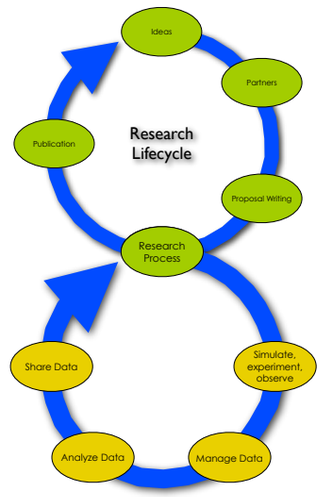


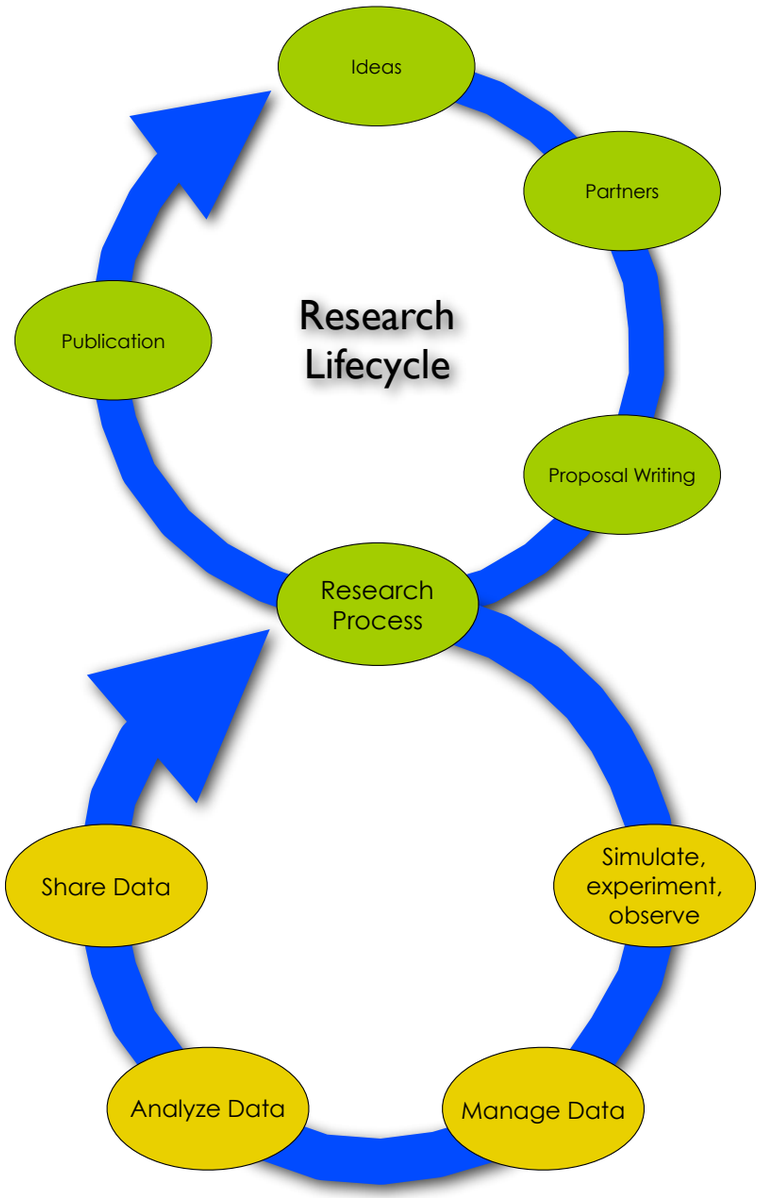
Research Lifecycle




Ideas
Partners
Proposal Writing
Publication
Research Lifecycle
Research Process
Share Data
Simulate, experiment, observe
Analyze Data
Manage Data

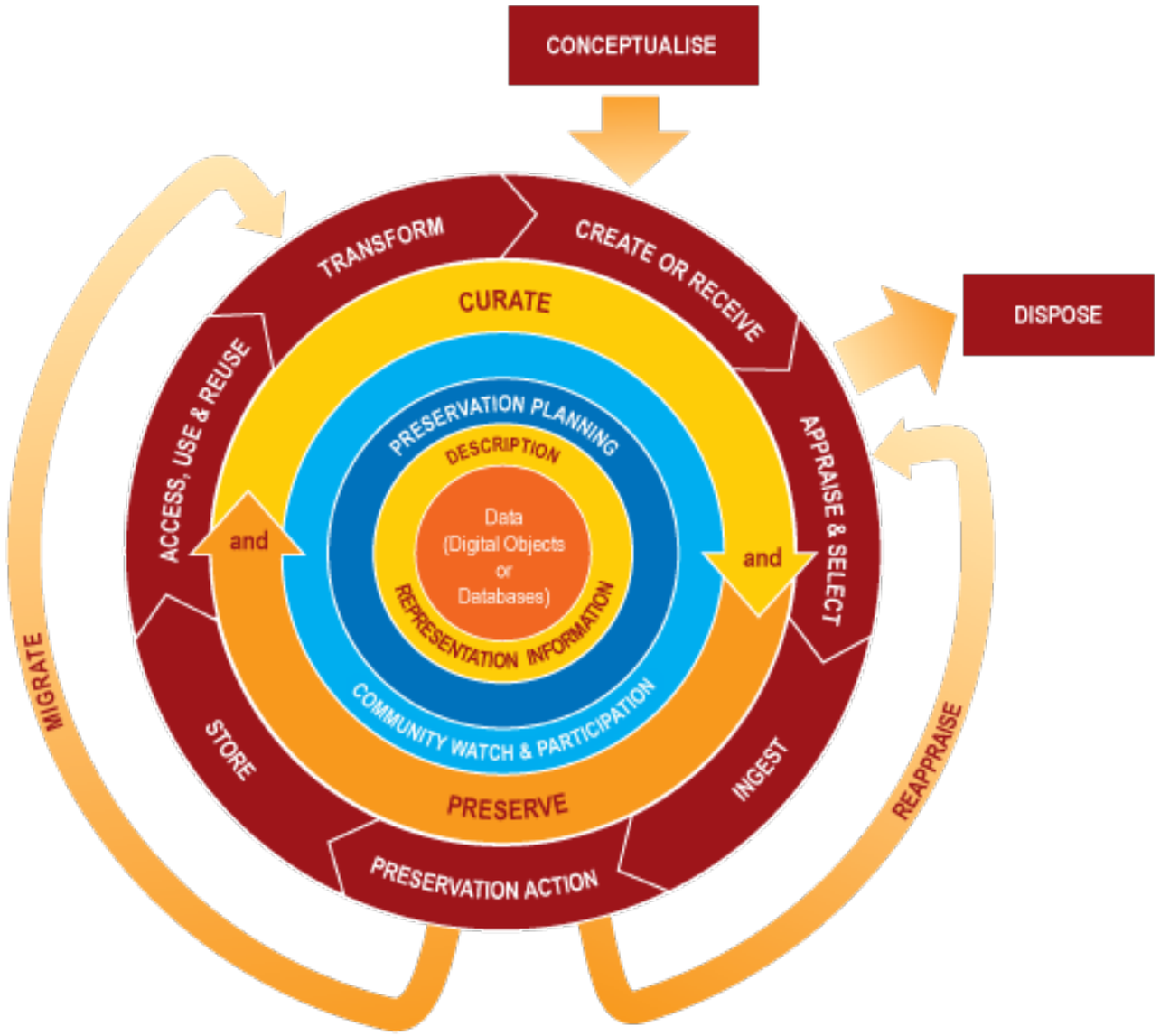

Ideas
Partners
Proposal Writing
Publication
Research Lifecycle
Research Process
Simulate, experiment, observe
Manage Data
Analyze Data
Share Data


Ideas
Partners
Proposal Writing
Research Process
Simulate, experiment, observe
Manage Data
Analyze Data
Share Data
Publication
Research Lifecycle



JISC Research Lifecycle

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



DCC Data Curation Lifecycle

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Research Lifecycle

- Ideas
- Partners
- Proposal Writing
- simulate, experiment, observe
- manage the data
- analyse data
- share data
- Publication

Research Process

Data Lifecycle

Curate | Preserve | Curate

Conceptualize	Create/ Receive	Appraise/ Select	Ingest	Preservation Action	Store	Access, Use, Reuse	Transform	Dispose

Reappraise

Migrate

“Data” and “Documentation”

A conversation we don't want to have

http://www.youtube.com/watch?v=N2zK3sAtr-4&feature=share&list=FLdHSa0dF8hj5B-x4OaIBPWQ

What were the assumptions of the original researcher? Were they reasonable?

What were the expectations of the new researcher? Were they reasonable?

What could have been done to ensure that the research data produced by the first researcher would be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR)

Some Clearinghouse Resources
FAIR Webinar Series

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

You are starting a new study, and you find a publication that is based on data key to your analysis …

Scenario 1

Hi
21

M. C
D. Th
C. O.

Quar
fores
fores
2000
exhib
well-
Mala
subtr
due l
These

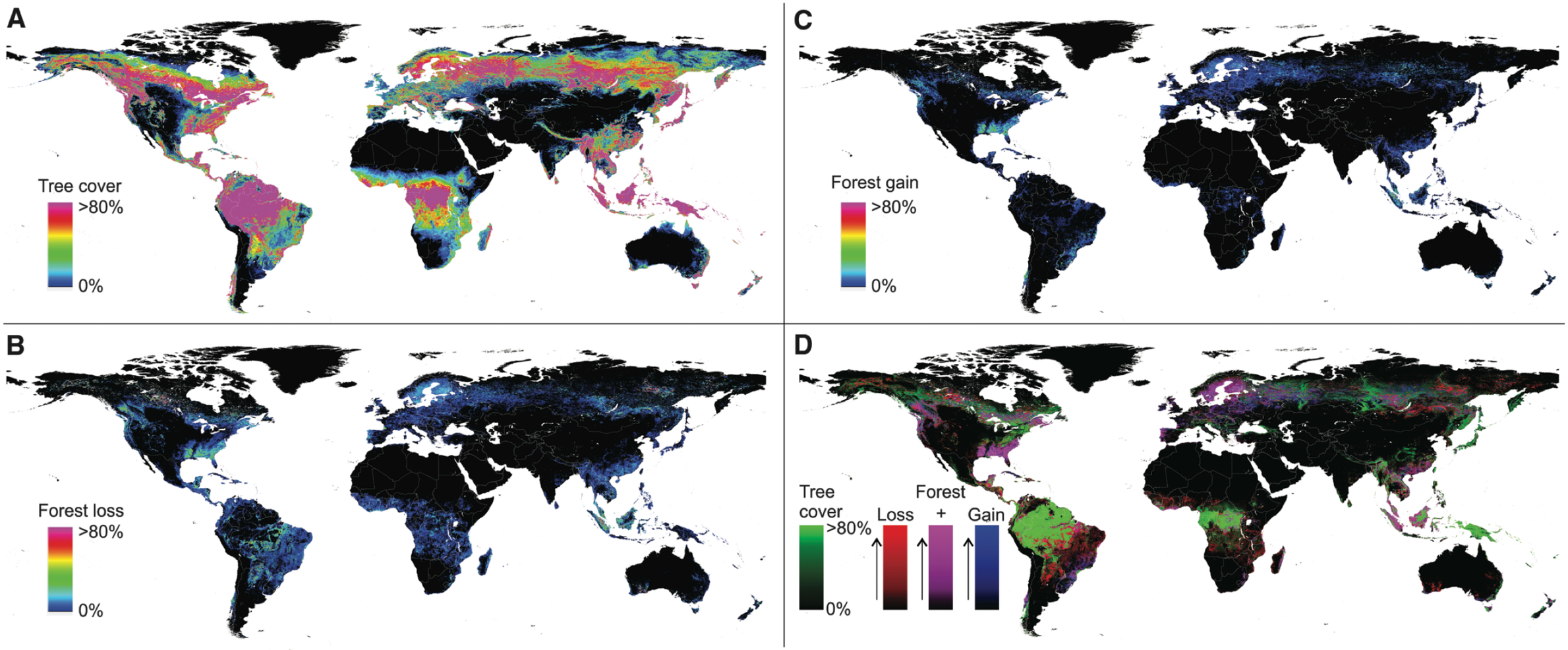


Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain. A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in (**D**), with loss and gain enhanced for improved visualization. All map layers have been resampled for display purposes from the 30-m observation scale to a 0.05° geographic grid.

High-Resolution Global Maps of 21st-Century Forest Cover Change

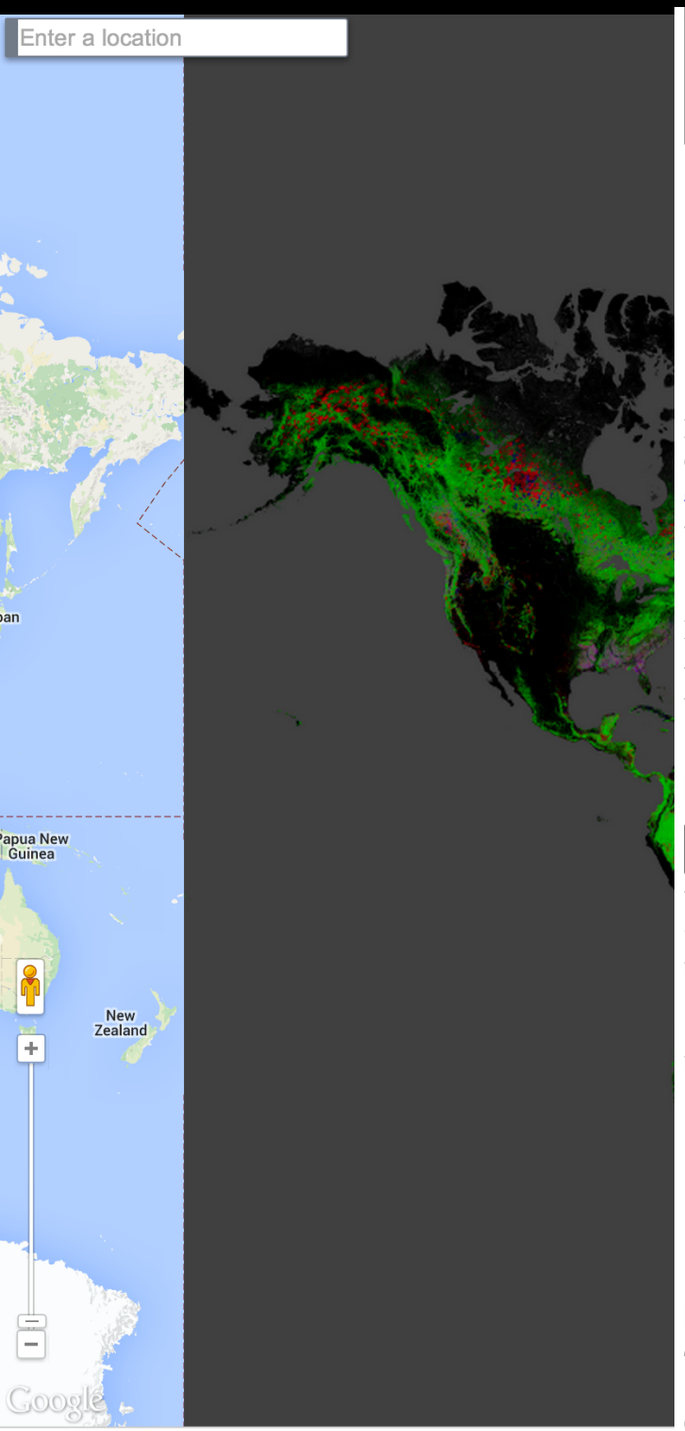
M. C. Hansen *et al.*
Science **342**, 850 (2013);
DOI: 10.1126/science.1244693

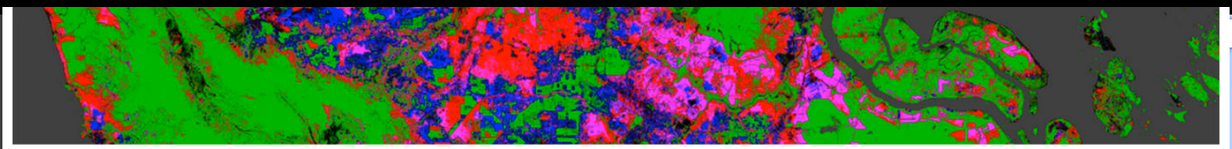
Questions …

Rate from 1 (impossible) to 5 (easy) the following

	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

Scenario 2





UNIVERSITY OF MARYLAND
DEPARTMENT OF GEOGRAPHICAL SCIENCES

Global Forest Change 2000–2012
Data Download

Results from time-series analysis of 654,178 Landsat 7 ETM+ images in characterizing global forest extent and change from 2000 through 2012. For additional information about these results, please see the associated journal article (Hansen et al., *Science* 2013).

Web-based visualizations of these results are also available at our main site:

http://earthenginepartners.appspot.com/science-2013-global-forest

Please use that URL when linking to this dataset.

We anticipate releasing updated versions of this dataset. To keep up to date with the latest updates, and to help us better understand how these data are used, please **register as a user**. Thanks!

License and Attribution

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy and redistribute the material in any medium or format, and to transform and build upon the material for any purpose, even commercially. You must give appropriate credit, provide a link to the license, and indicate if changes were made.

Use the following credit when these data are displayed:

Source: Hansen/UMD/Google/USGS/NASA

Use the following credit when these data are cited:

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850–53. Data available on-line from: http://earthenginepartners.appspot.com/science-2013-global-forest.

Dataset Details

This global dataset is divided into 10x10 degree tiles, consisting of seven files per tile. All files contain unsigned 8-bit values and have a spatial resolution of 1 arc-second per pixel, or approximately 30 meters per pixel at the equator.

Published by Ha

Global Forest Change

Published by Hansen, Potapov, Moore, Hancher et al.

UNIVERSITY OF MARYLAND
DEPARTMENT OF GEOGRAPHICAL SCIENCES

Results from time-series analysis of 654,178 Landsat images in characterizing forest extent and change, 2000–2012.

Trees are defined as all vegetation taller than 5m in height and are expressed as a percentage per output grid cell as '2000 Percent Tree Cover'. 'Forest Loss' is defined as a stand-replacement disturbance, or a change from a forest to non-forest state. 'Forest Gain' is defined as the inverse of loss, or a non-forest to forest change entirely within the study period. 'Forest Loss Year' is a disaggregation of total 'Forest Loss' to annual time scales.

Reference 2000 and 2012 imagery are median observations from a set of quality assessment-passed growing season observations.

Download the data.

Reset to default view

Data Products

Loss/Extent/Gain (Red/Green/Blue)

Legend
- Forest Loss 2000–2012
- Forest Gain 2000–2012
- Both Loss and Gain
- Forest Extent

Background Imagery

Year 2000 Bands 5/4/3

Example Locations

Forestry and Tornado in Alabama

Zoom to area

The trail of destruction from the April 27 2011 Tuscaloosa-Birmingham tornado is clearly visible in this location. This was one of 358 recorded tornadoes during the April 25-28, 2011 tornado outbreak, the most severe in US history.

Zoom out to spot tracks from other tornadoes nearby.

http://earthenginepartners.appspot.com/science-2013-global-forest.html

Questions …

Rate from 1 (impossible) to 5 (easy) the following

	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

discussion

Some Definitions

Data

Data. For the purposes of this document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.

- National Science Foundation (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Cyberinfrastructure Council. Washington, DC. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf. pg. 22

Documentation (AKA Metadata)

Metadata. Metadata are a subset of data, and are data about data. Metadata summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.

- National Science Foundation (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Cyberinfrastructure Council. Washington, DC. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf. pg. 22

Embargo

Embargo. A period during which access to research data is not allowed to certain types of users. This is either to protect the revenue of the publisher or (more generally) to protect the interests of other parties (for example, partner research organizations).

— University of Bristol: Research Data Management Glossary

License

A license in this context is a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held.The first thing to note is that only the rights holder (or someone with a right or license to act on their behalf) can grant a license; it is therefore imperative that the intellectual property rights (IPR) pertaining to the data are established before any licensing takes place.

How to License Research Data. Digital Curation Centre.
http://www.dcc.ac.uk/resources/how-guides/license-research-data#x1-20002

Some Recommendations

I'M EXHAUSTED FROM ALL OF THE BASIC RESEARCH I'M DOING.

IT'S TOO BAD THAT THE VALUE OF MY WORK WON'T BE QUANTIFIABLE FOR ANOTHER TEN YEARS.

I'D LIKE TO SEE YOUR LAB REPORT.

SO... THE NEW RULE IS THAT WE WRITE DOWN STUFF?

What you need to know …

WHO?
WHAT?
WHERE?
WHEN?
WHY?
HOW?
ACCESS?

- **WHO?** CREDIT (RESEARCHERS, SPONSORS), QUESTIONS, RESPONSIBILITY, ROLE
- **WHAT?** WHAT WAS MEASURED, UNITS, AGGREGATION
- **WHERE?** GEOGRAPHIC LOCATION (DEFINE DATUM, COORDINATE SYSTEM, METHOD)
- **WHEN?** DATE, TIME - STRUCTURED, CONSISTENT, TIME ZONE, STANDARDS-BASED
- **WHY?** PURPOSE FOR DATA COLLECTION, SUGGESTED USE, KNOWN LIMITATIONS
- **HOW?** INSTRUMENTS, SENSORS, ALGORITHMS, MODELS, SOFTWARE
- **ACCESS?** LICENSING TERMS, EMBARGO, REDISTRIBUTION, MODIFICATION

Organization

Define folder and file names and structure - and use it

Use meaningful names that include basic information (e.g. date, measurement, collection, etc.)

Unique

Avoid Spaces

ASCII Characters only

Some Clearinghouse Resources

Search for "organizing data" in the Clearinghouse and try out some Keyword facets

A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

http://www.phdcomics.com/comics/archive.php?comicid=1323

Structure/Content

Consistent content

Separate data from analysis

Keep raw data separate

Focus on tabular structure for tabular data

Explicitly encode missing data, and document that encoding

Use meaningful column headings - while keeping short without spaces

Include units

Data dictionary

Some Clearinghouse Resources

Go to Clearinghouse and search for "tabular" and/or "spreadsheets"

Formats

Plan for integration into an archive

Open Standards

>

Proprietary ASCII

>

Proprietary Binary - Documented

>

Proprietary Binary

Some Clearinghouse Resources

Go to Clearinghouse and search for "data formats"

and then try some keyword facets

Documentation

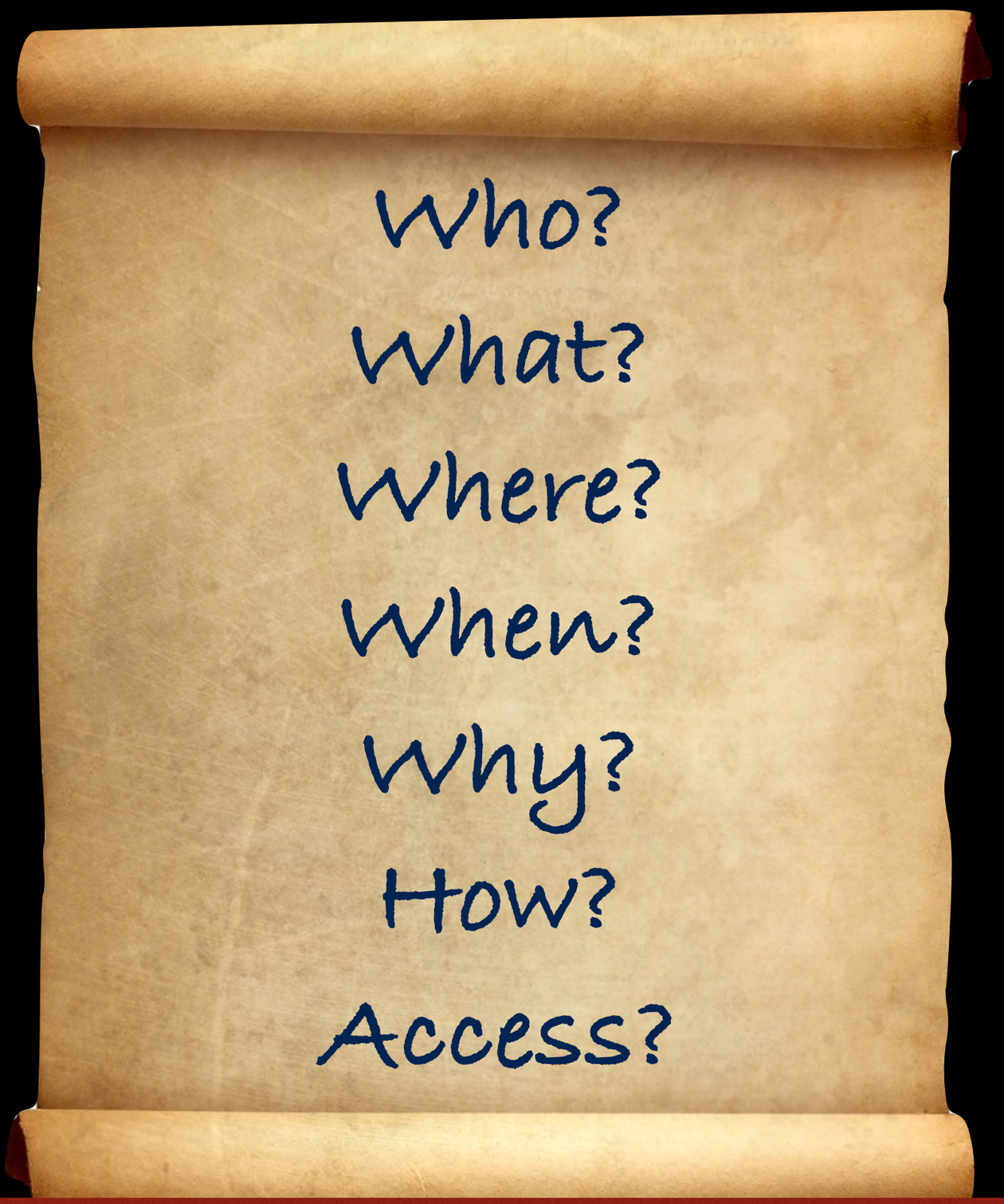
Many documentation standards

Machine and human readable

Commonly based on Extensible Markup Language (XML)

Wide variety of strategies/ methods/tools for creating documentation

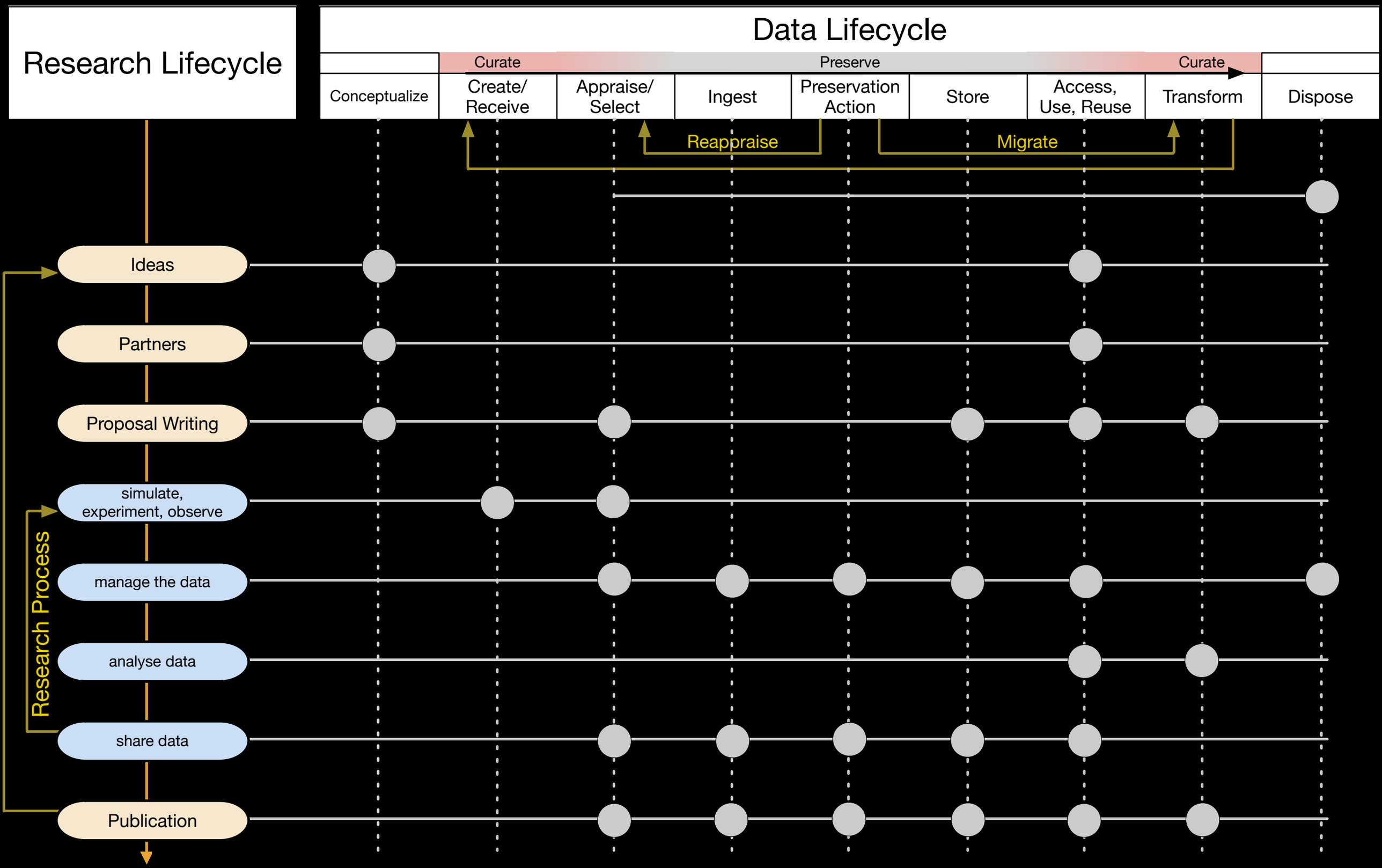
Enables *Discovery*, *Use*, and *Understanding*



Some Clearinghouse Resources

Search for "metadata" and /or "documentation" in the Clearinghouse and then try out some keyword facets

A Process


Research Lifecycle
Data Lifecycle
Curate
Preserve
Curate
Conceptualize
Create/ Receive
Appraise/ Select
Ingest
Preservation Action
Store
Access, Use, Reuse
Transform
Dispose
Reappraise
Migrate
Ideas
Partners
Proposal Writing
simulate, experiment, observe
manage the data
analyse data
share data
Publication
Research Process

Resources

- Workshop Presenters:
  Karl Benedict - kbene@unm.edu
  Nancy Hoebelheinrich - nhoebel@kmotifs.com

- Library of Congress: Sustainability of Digital Formats:
  www.digitalpreservation.gov/formats/index.shtml

- Digital Curation Centre: Disciplinary Metadata:
  www.dcc.ac.uk/resources/metadata-standards

- ESIP Data Management Training Clearinghouse
  dmtclearinghouse.esipfed.org

- re3data.org - Registry of Research Data Repositories
  www.re3data.org

- Repository Finder tool:
  https://repositoryfinder.datacite.org.

- Creative Commons
  creativecommons.org/share-your-work/

Acknowledgements

- Unless otherwise noted, all images are from shutterstock.com
- The work upon which this workshop is based has been sponsored by the
    - The U.S. Institute of Museum and Library Services (IMLS Grant LG-70-18-0092-18)
    - NSF EPSCoR Program (Track 1 [Awards: 0447691, 0814449, 1301346] and Track 2 awards [0918635, 1329470])
    - New Mexico Resource Geographic Information System
    - NASA ACCESS Program
    - University of New Mexico - University Libraries
- The *Data Management Training Clearinghouse* is managed by the Earth Science Information Partners and was originally developed through a collaboration between ESIP, DataONE, and the USGS Community for Data Integration.

Questions?

Sign-In & Feedback: https://www.surveymonkey.com/r/rds-signin