Research Data Management - Introduction

Karl Benedict
Director of Research Data Services & Associate Professor, College of University Libraries & Learning Sciences
University of New Mexico

kbene@unm.edu

Outline

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

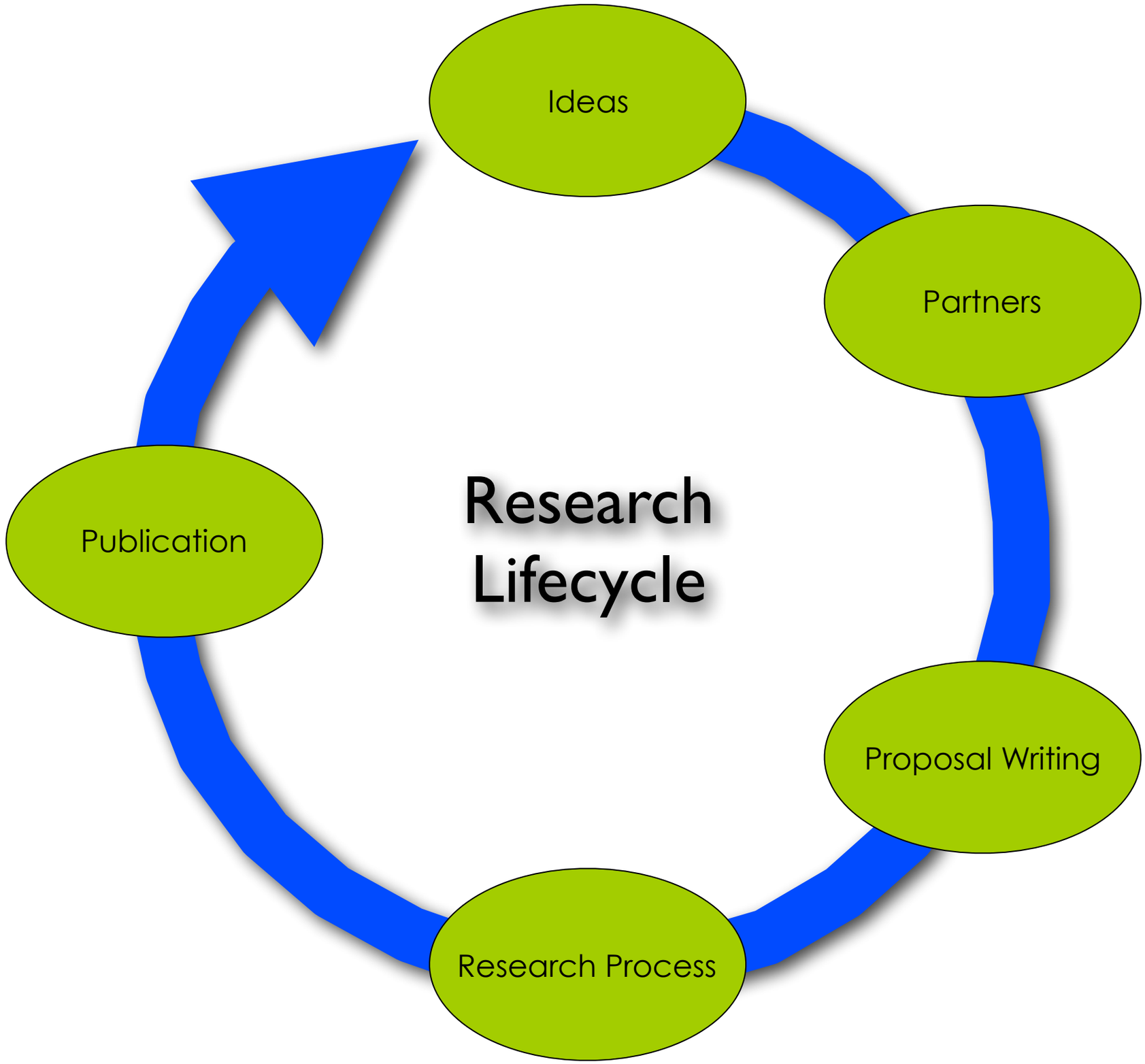
Context

Data Management Requirements

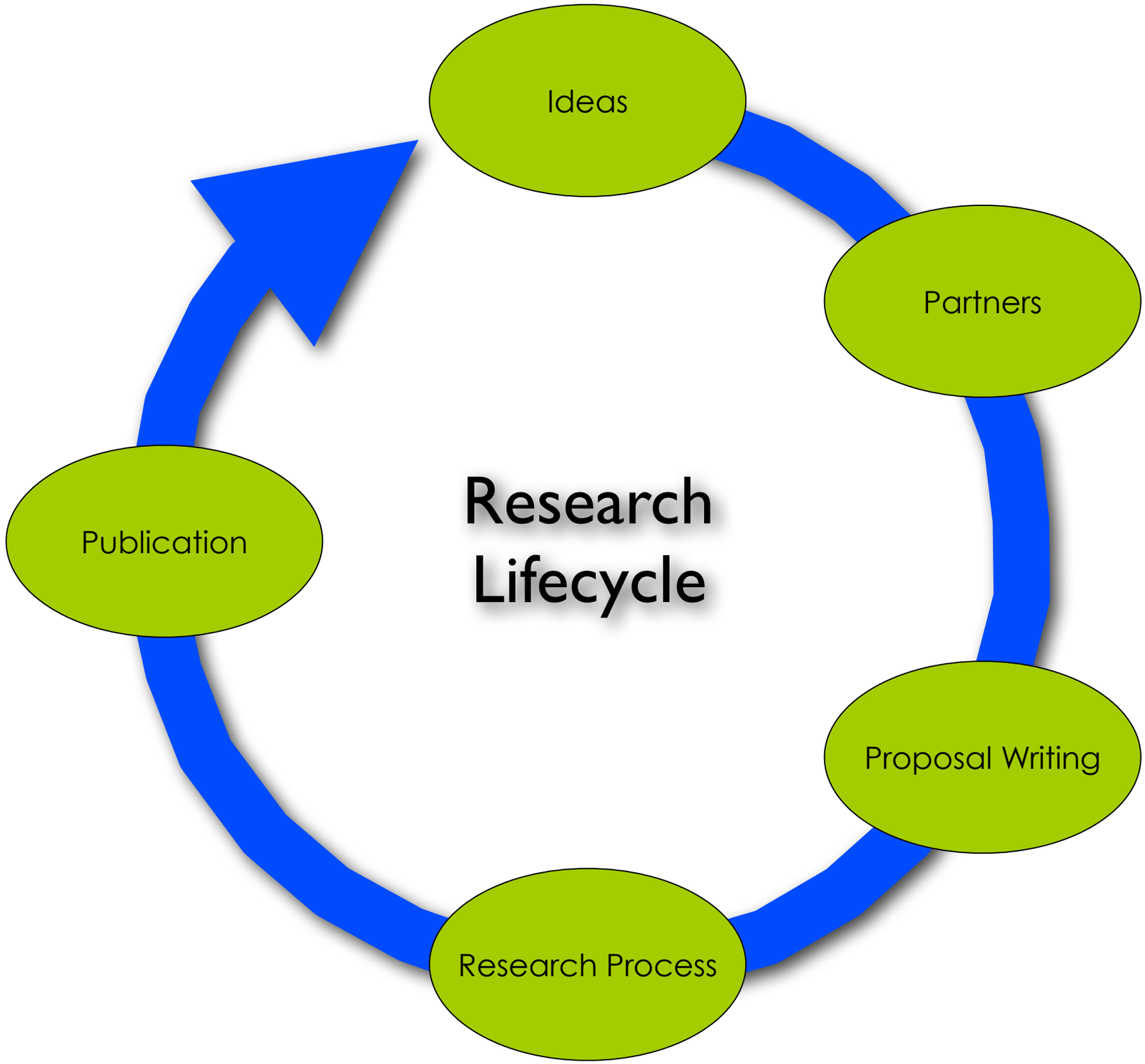
- Data Management Plans required by funding agencies
- Publications requiring links to supporting data
- Increasing collaborative research where data must be shared

A Cross-walk between the Research and Data Lifecycles

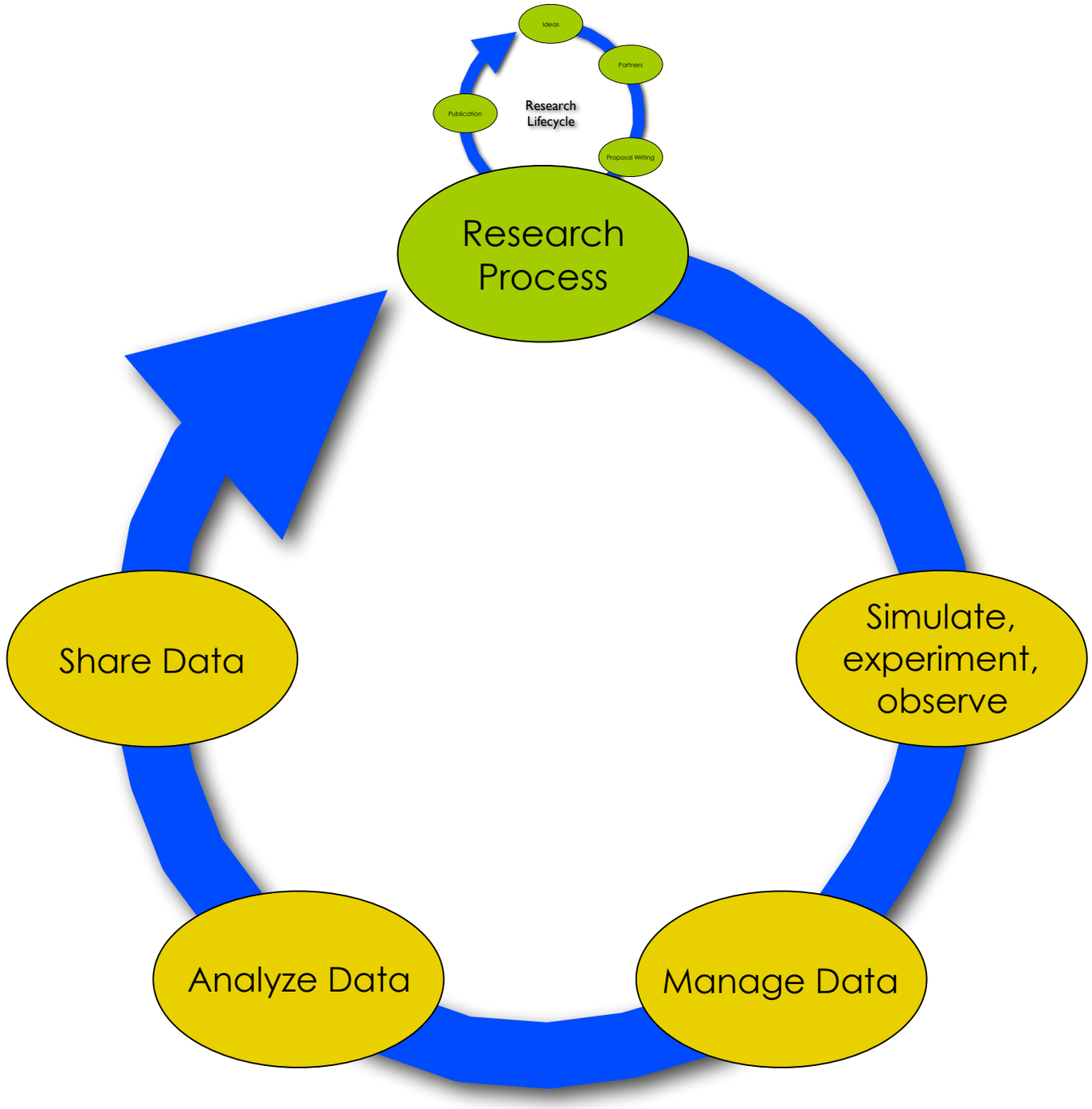
Research Lifecycle

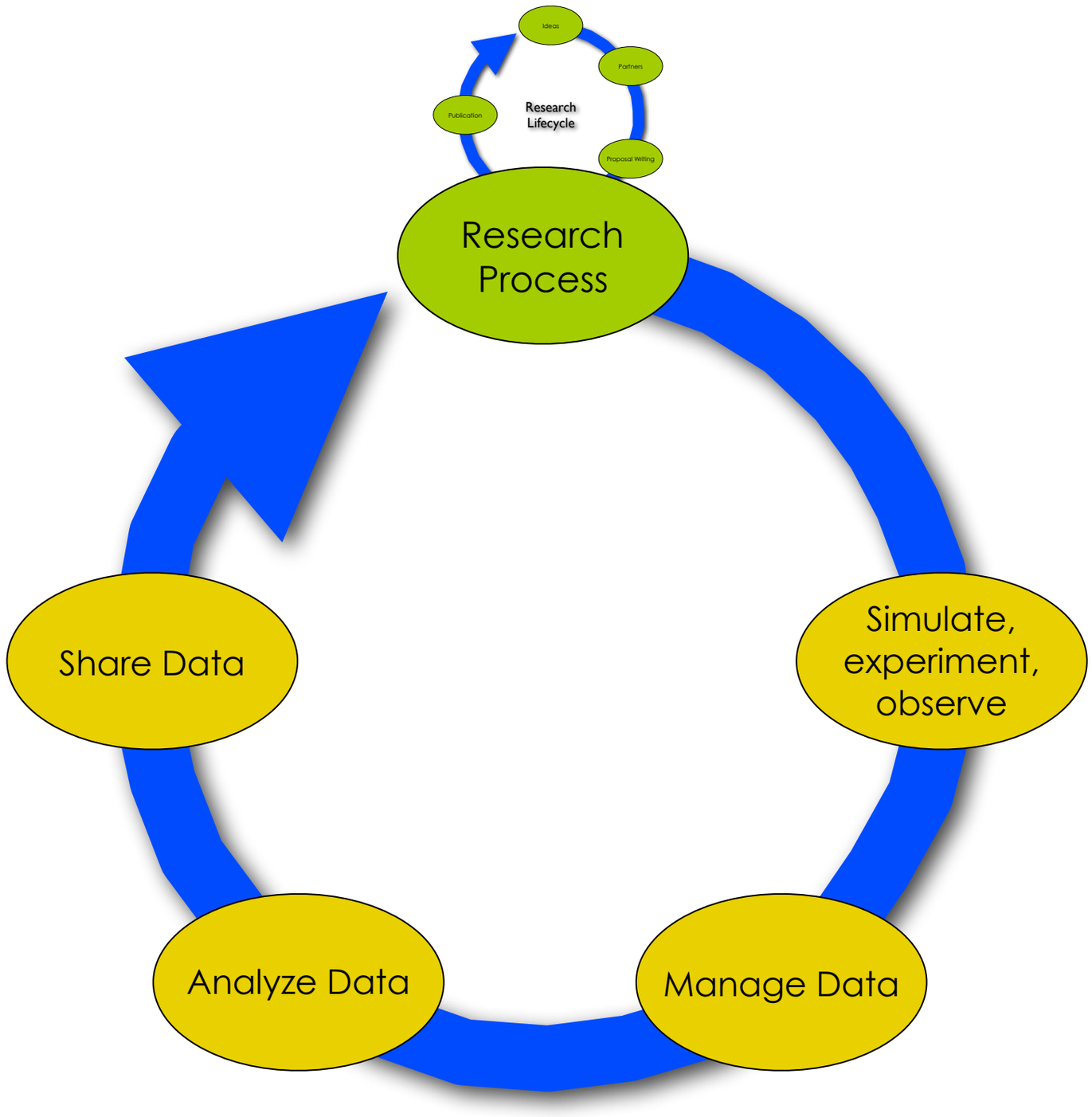


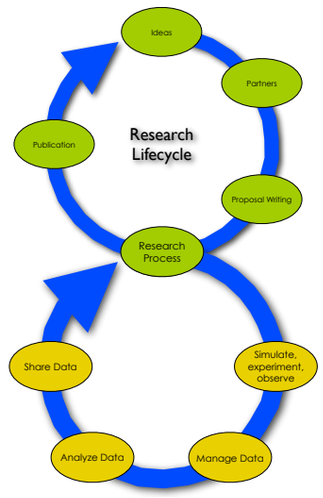
Research Lifecycle

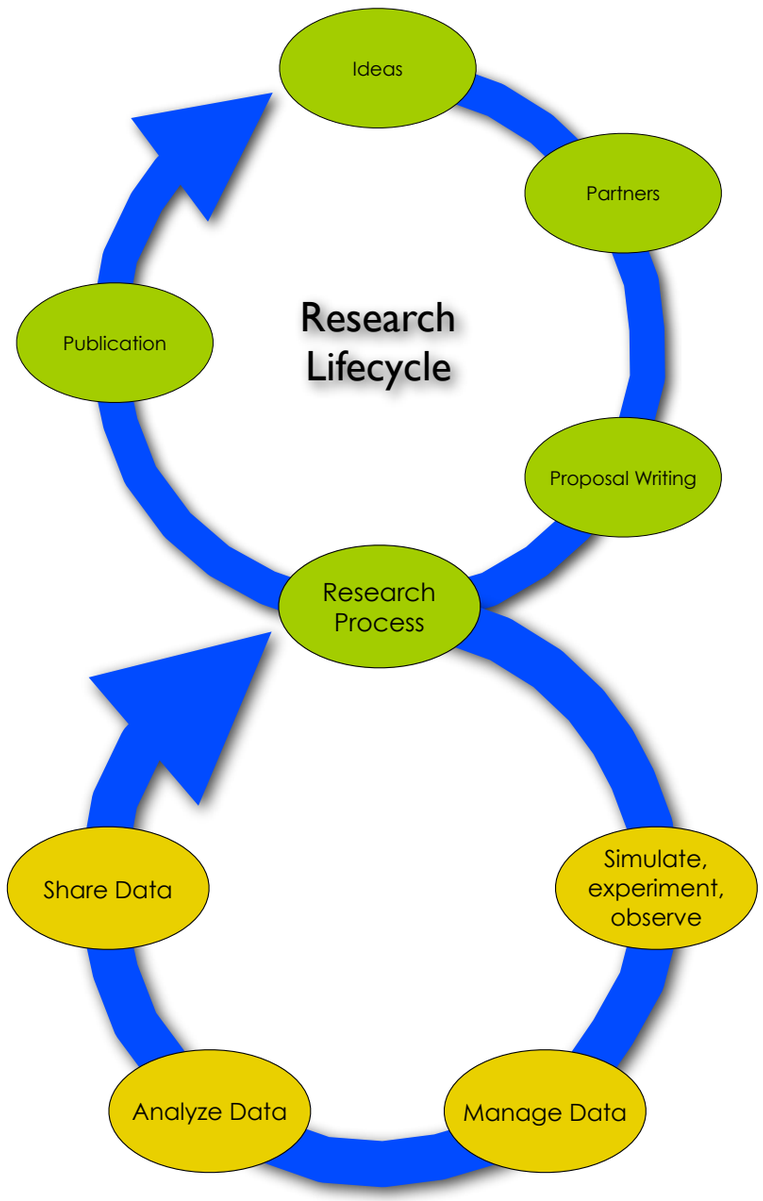


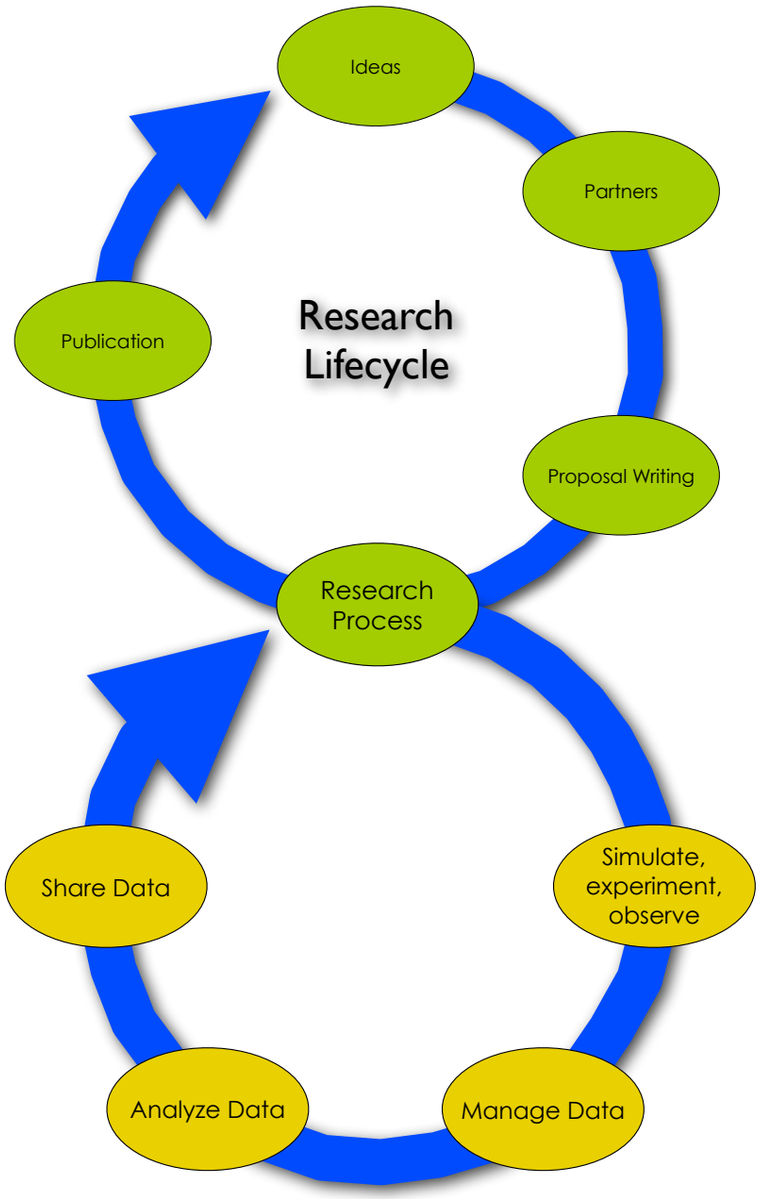

Ideas
Partners
Proposal Writing
Research Process
Publication
Research
Lifecycle


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

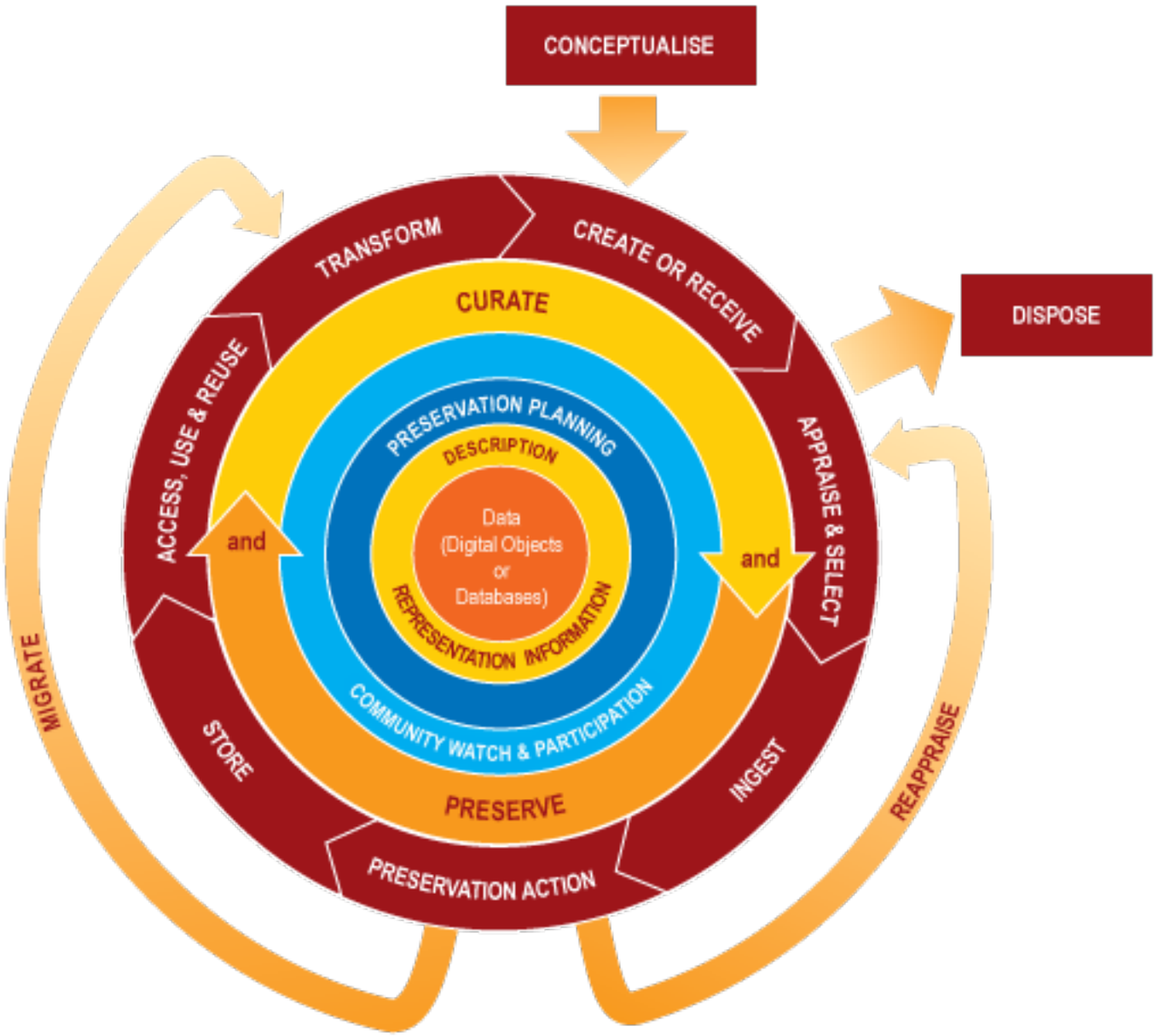

Ideas
Partners
Proposal Writing
Research Process
Simulate, experiment, observe
Manage Data
Analyze Data
Share Data
Publication
Research Lifecycle


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



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

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

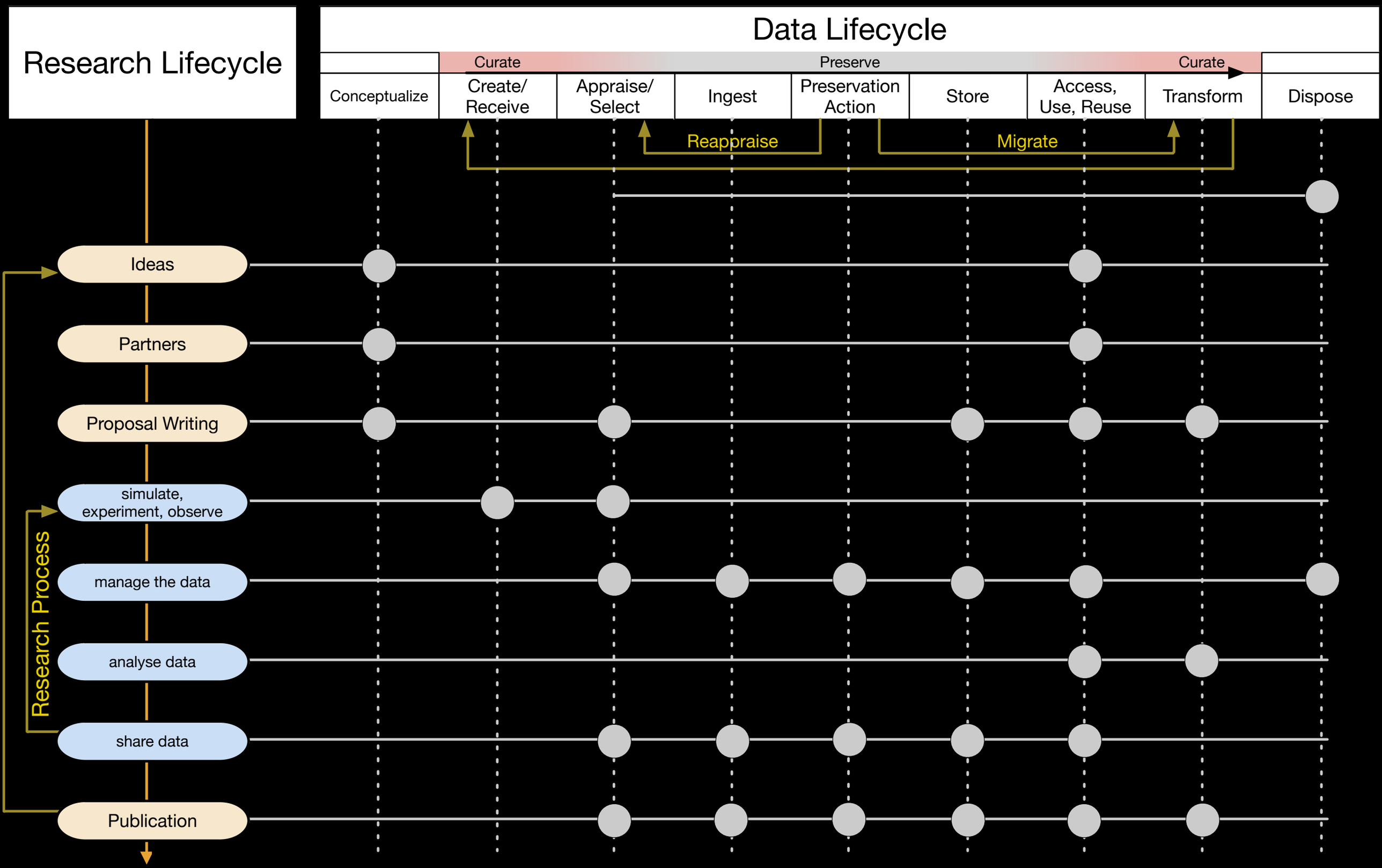
manage the data

analyse data

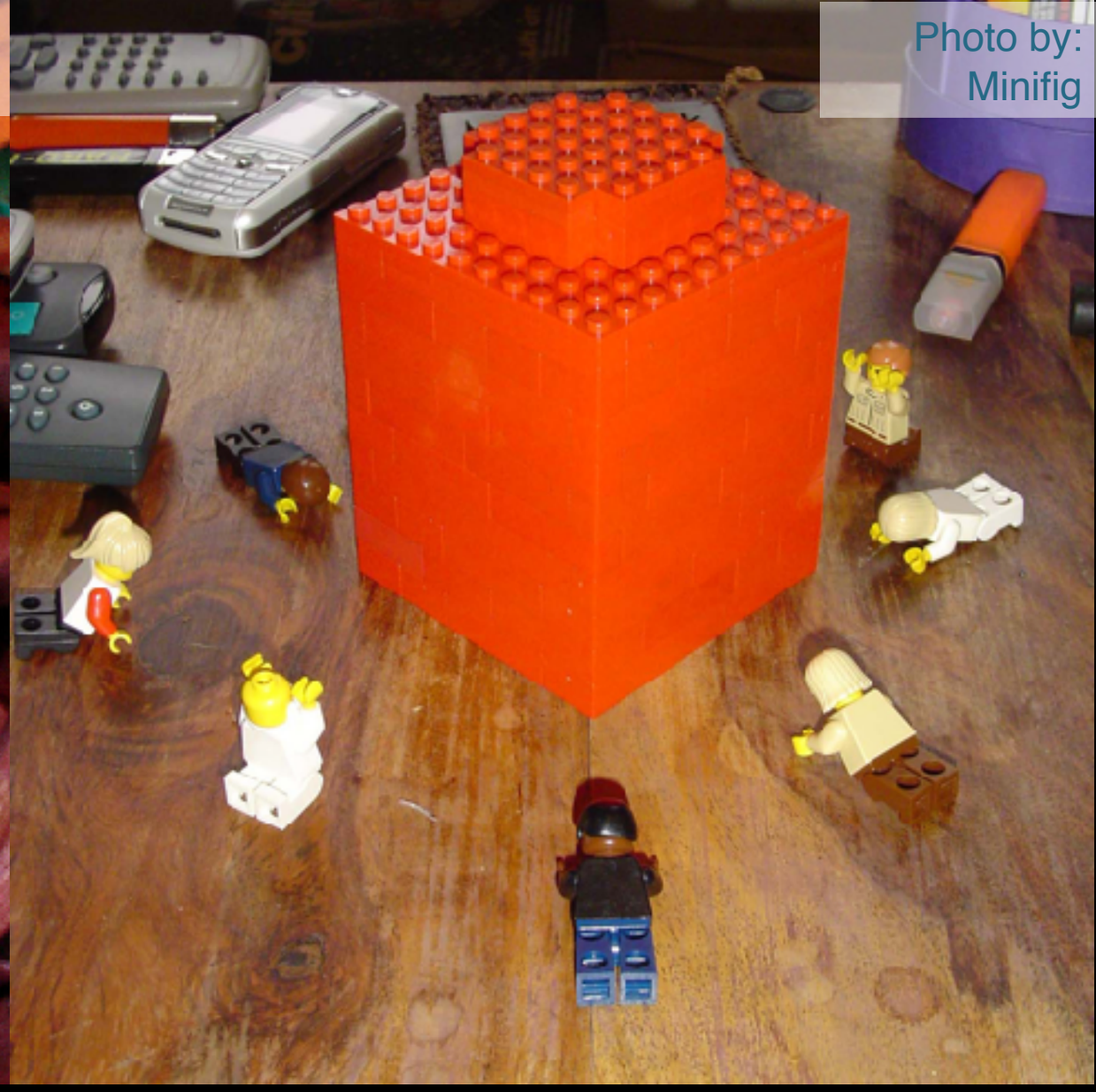
share data

Publication

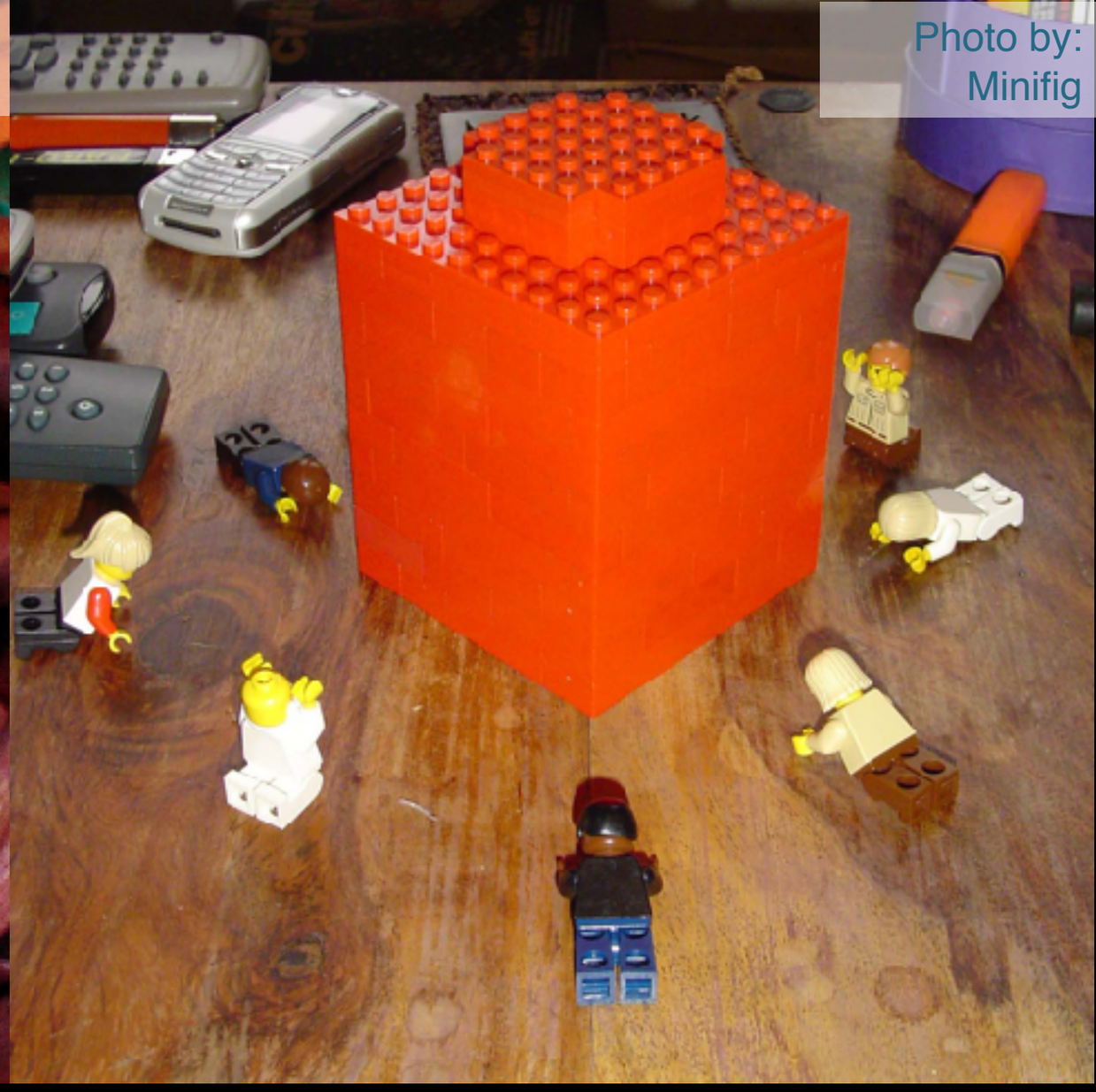
Research Process


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

Building



10 minutes
Choose 15 pieces
Place a being at the top
Build for height (*impact)*
Build for stability (*support)*



10 minutes
Choose 15 pieces
Place a being at the top
Build for height (*impact)*
Build for stability (*support)*

Who “Won”?

Why?

“Data” and “Documentation”

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

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

Scenario 1

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

M. C. Hansen,[1*] P. V. Potapov,[1] R. Moore,[2] M. Hancher,[2] S. A. Turubanova,[1] A. Tyukavina,[1] D. Thau,[2] S. V. Stehman,[3] S. J. Goetz,[4] T. R. Loveland,[5] A. Kommareddy,[6] A. Egorov,[6] L. Chini,[1] C. O. Justice,[1] J. R. G. Townshend[1]

Quantification of global forest change has been lacking despite the recognized importance of forest ecosystem services. In this study, Earth observation satellite data were used to map global forest loss (2.3 million square kilometers) and gain (0.8 million square kilometers) from 2000 to 2012 at a spatial resolution of 30 meters. The tropics were the only climate domain to exhibit a trend, with forest loss increasing by 2101 square kilometers per year. Brazil's well-documented reduction in deforestation was offset by increasing forest loss in Indonesia, Malaysia, Paraguay, Bolivia, Zambia, Angola, and elsewhere. Intensive forestry practiced within subtropical forests resulted in the highest rates of forest change globally. Boreal forest loss due largely to fire and forestry was second to that in the tropics in absolute and proportional terms. These results depict a globally consistent and locally relevant record of forest change.

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

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

Scenario 1

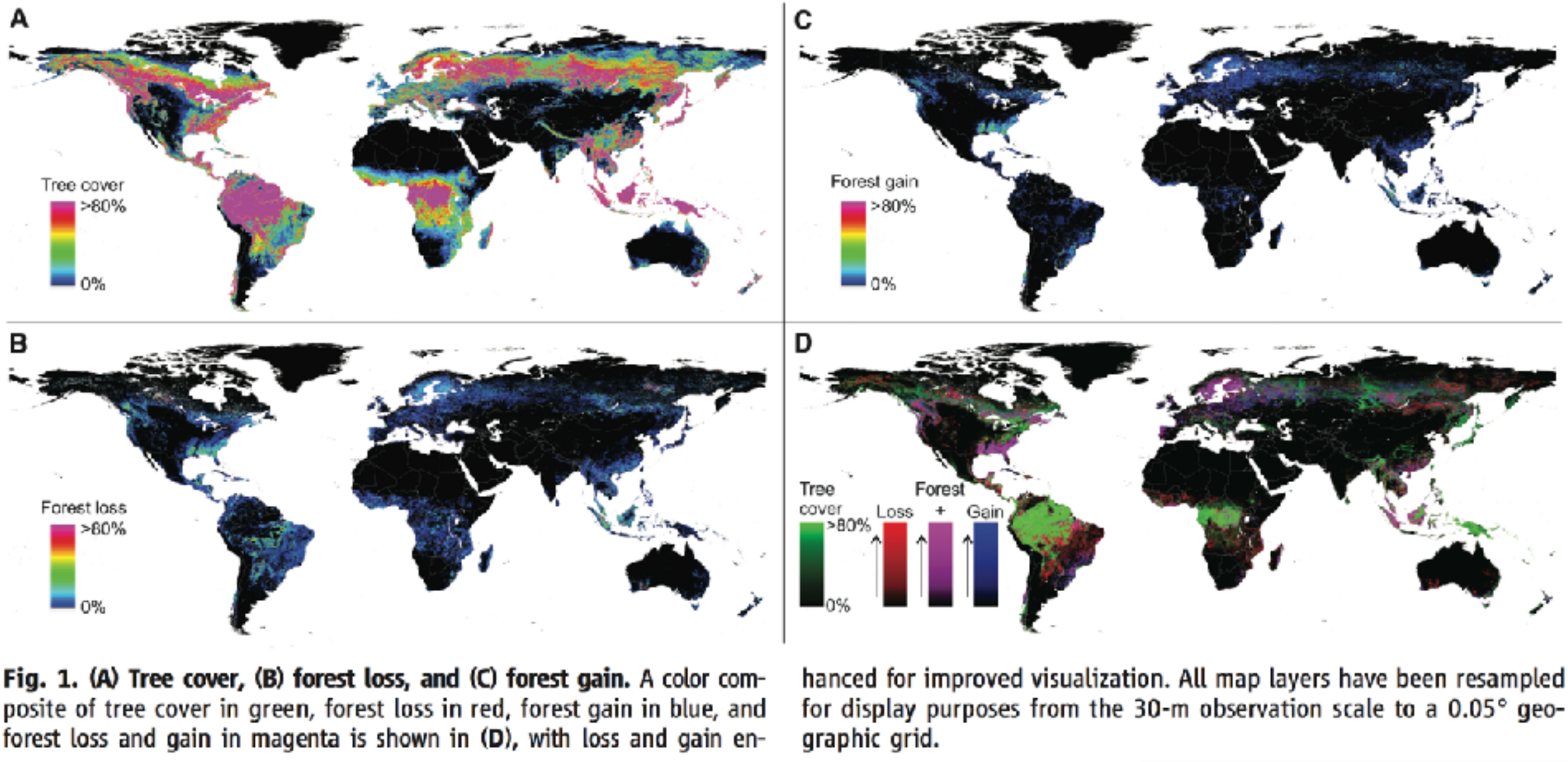


Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain. A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in **(D)**, with loss and gain enhanced for improved visualization. All map layers have been resampled for display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

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

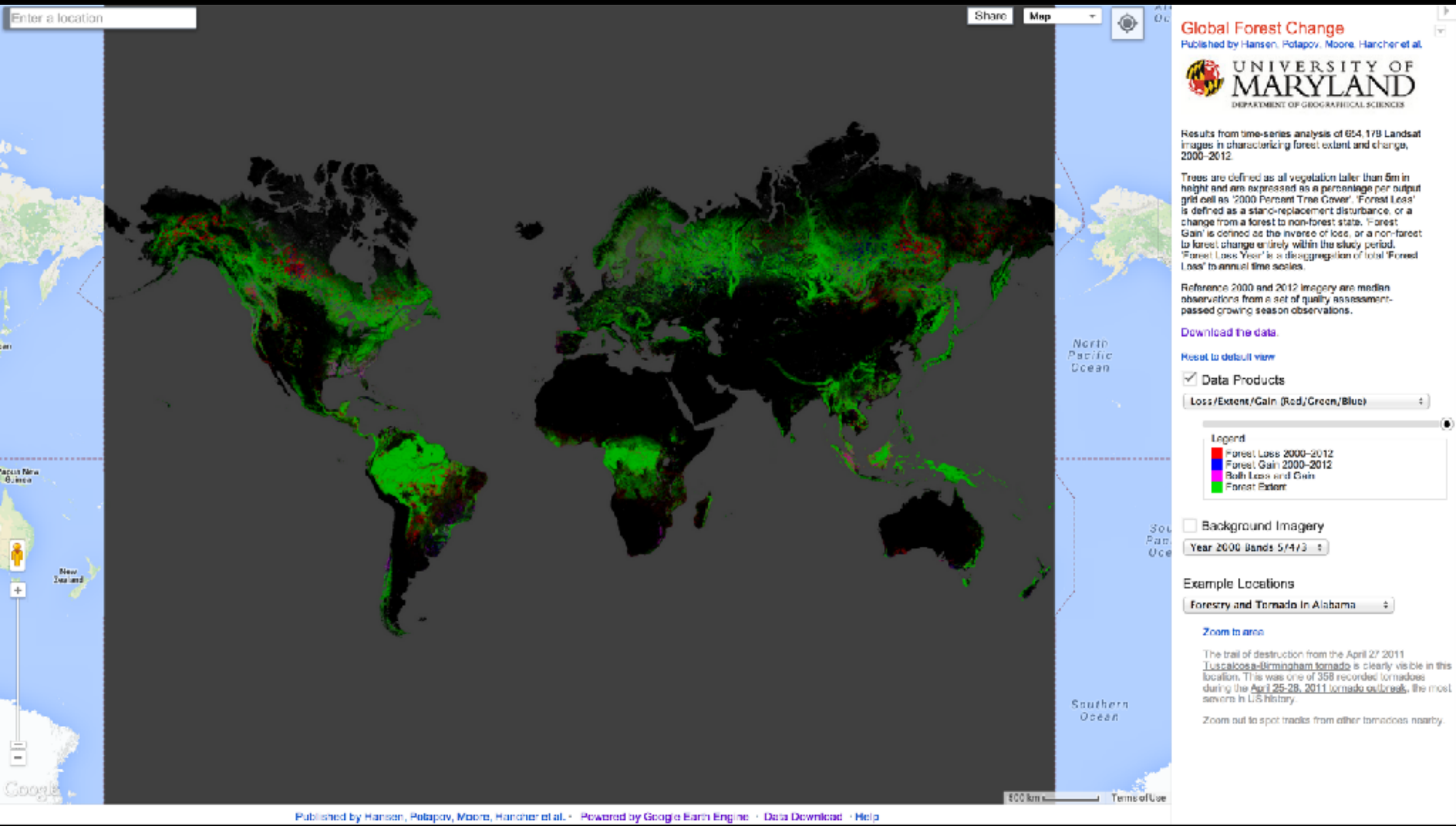
M. C. Hansen *et al.*
Science **342**, 850 (2013);
DOI: 10.1126/science.1244693

Questions …

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

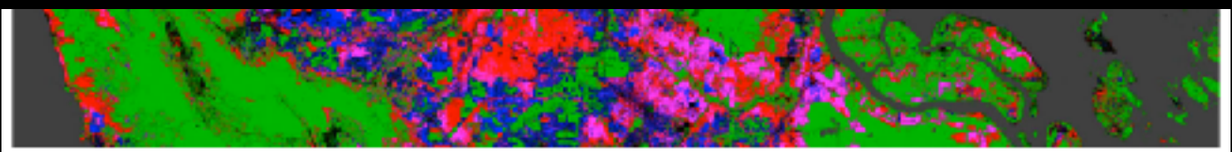
	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

Scenario 2



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

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.

http://earthenginepartners.appspot.com/science-2013-global-forest/download.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 licence 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 licence to act on their behalf) can grant a licence; 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

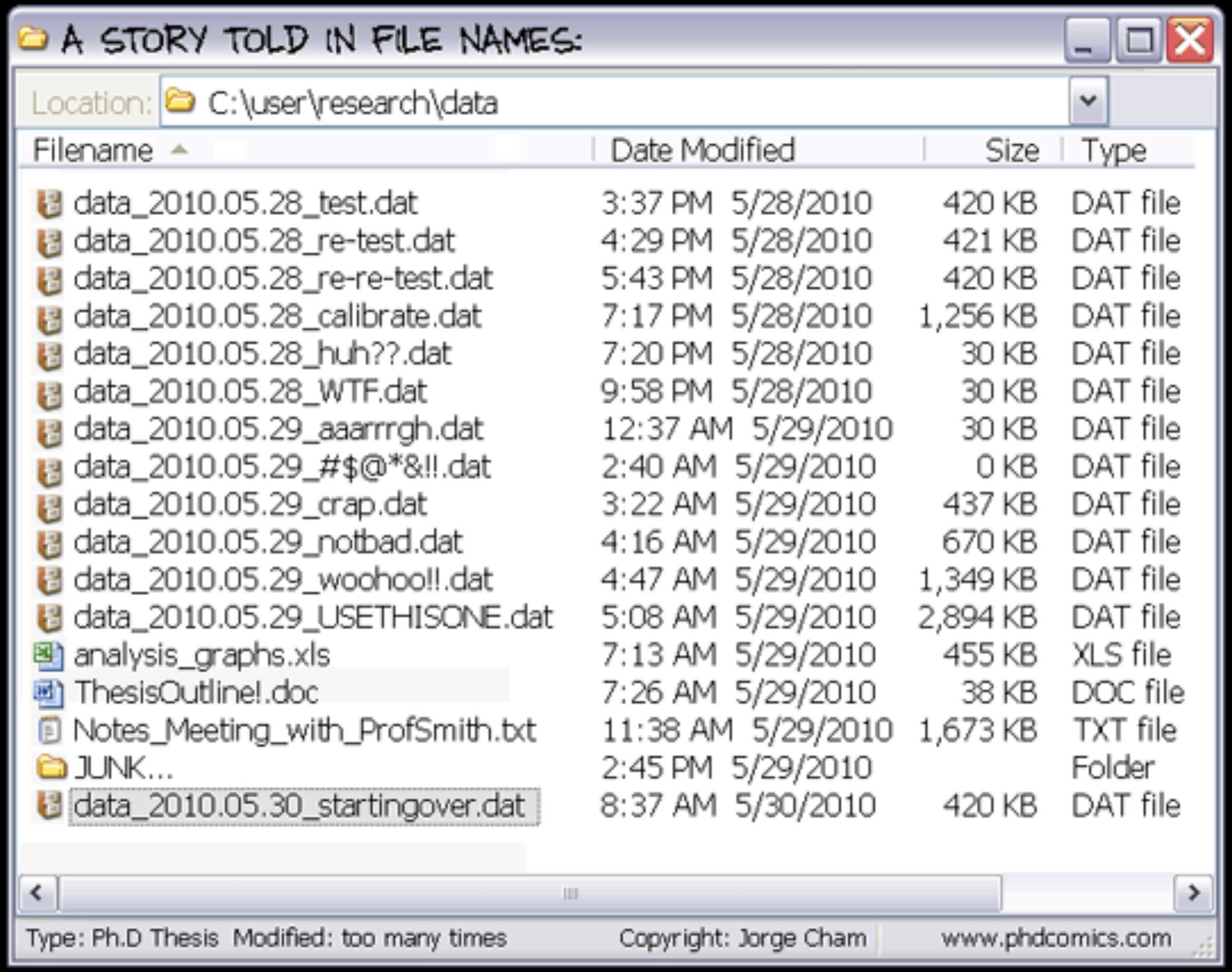
Define folder and file names and structure - and use it

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

Unique

Avoid Spaces

ASCII Characters only



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

Formats

Plan for integration into an archive

Open Standards
>
Proprietary ASCII
>
Proprietary Binary - Documented
>
Proprietary Binary

Documentation

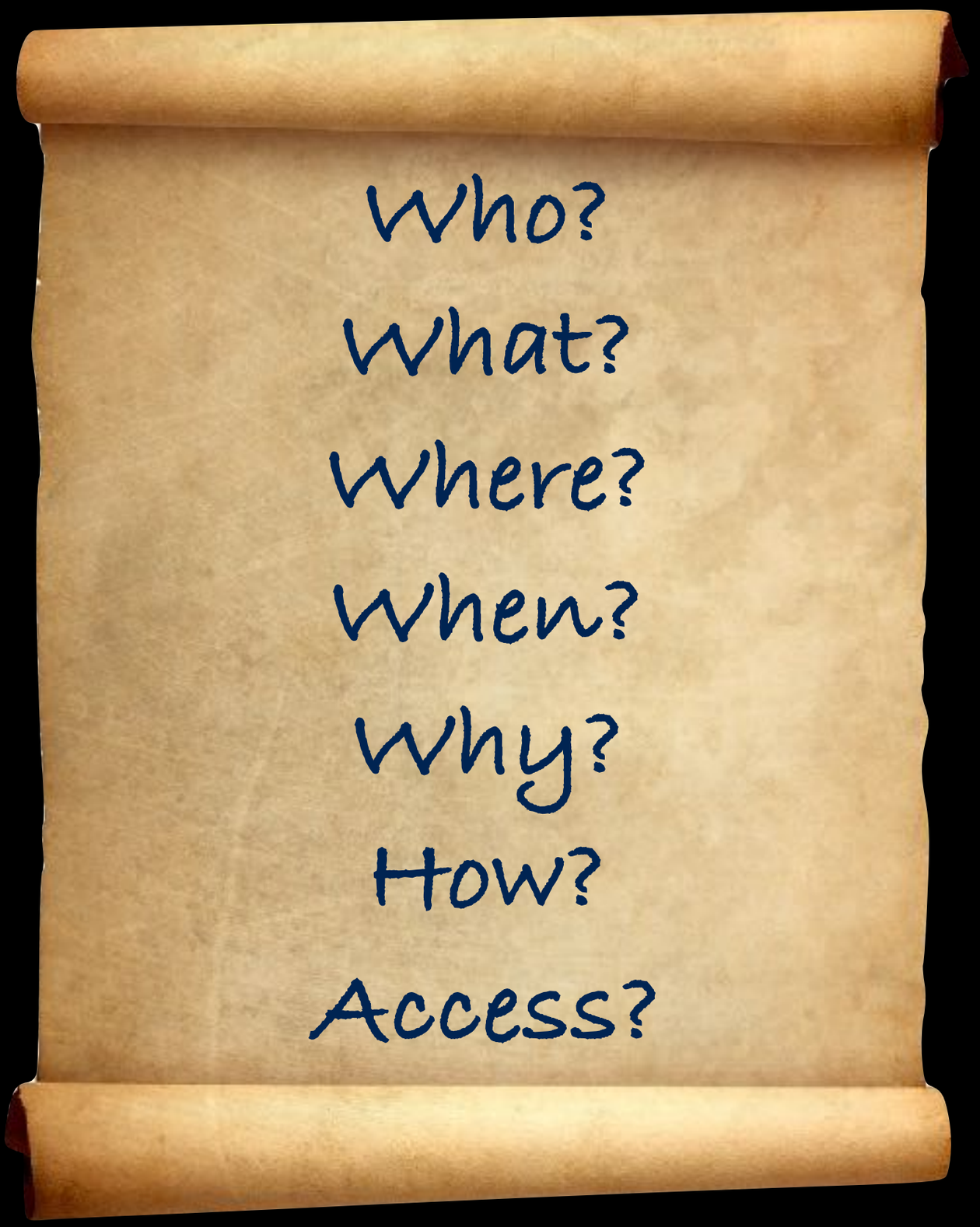
Many documentation standards

Machine and human readable

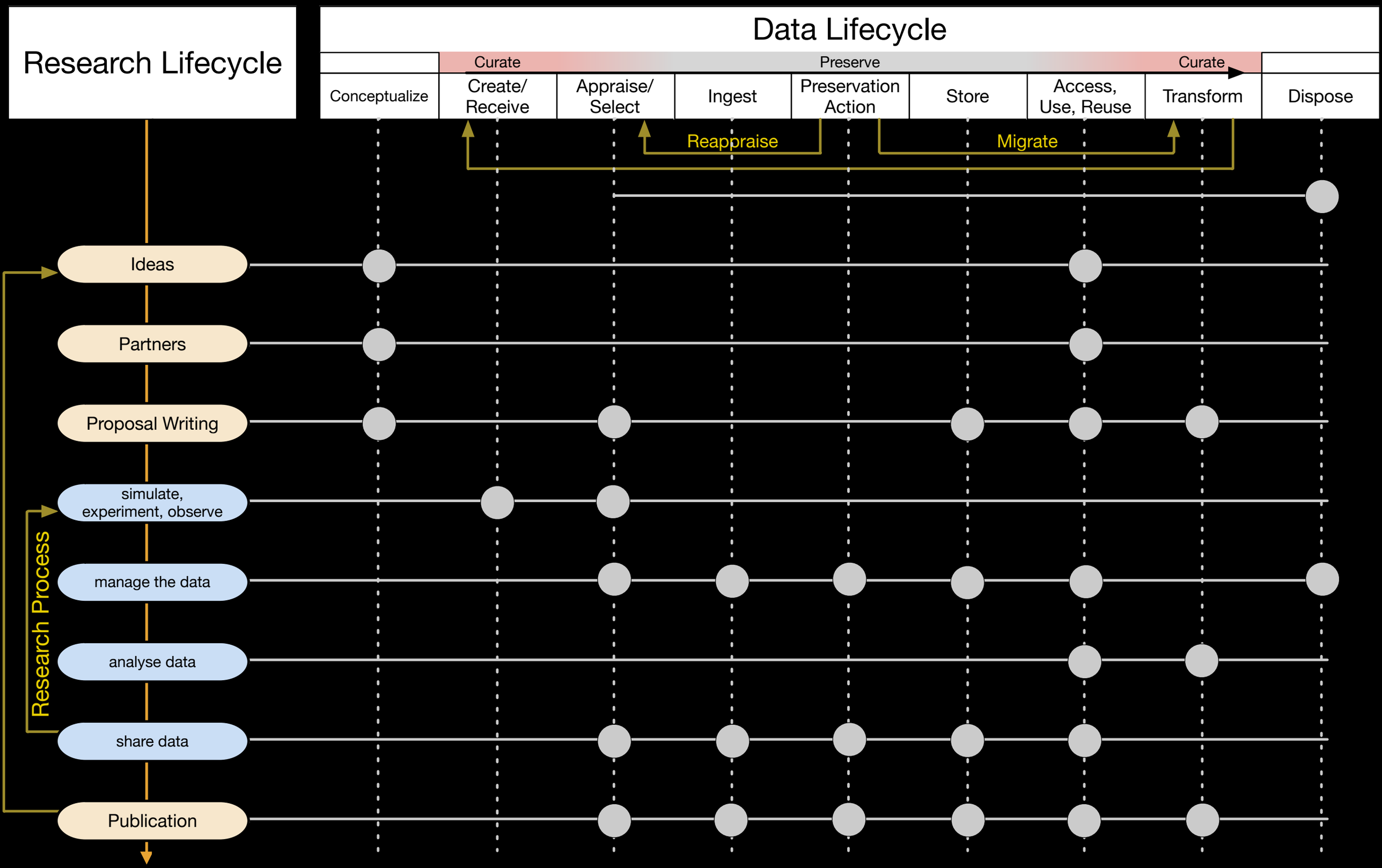
Commonly based on Extensible Markup Language (XML)

Wide variety of strategies/ methods/tools for creating documentation

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



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

- UNM Libraries - Research Data Services:
  Karl Benedict - kbene@unm.edu
  Jon Wheeler - jwheel01@unm.edu
  http://libguides.unm.edu/data

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

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

- Library Hosted Infrastructure:
  Digital Repository: http://digitalrepository.unm.edu
  Lobo Drive: http://lobodrive.unm.edu
  LoboGit: http://lobogit.unm.edu

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

Acknowledgements

- Unless otherwise noted, all images are from shutterstock.com
- The products and platforms upon which this work is based have been funded by
  - 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

Questions?

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