Committing to Data Quality

Ann Green
Digital Lifecycle Research & Consulting

NADDI
Vancouver 2014

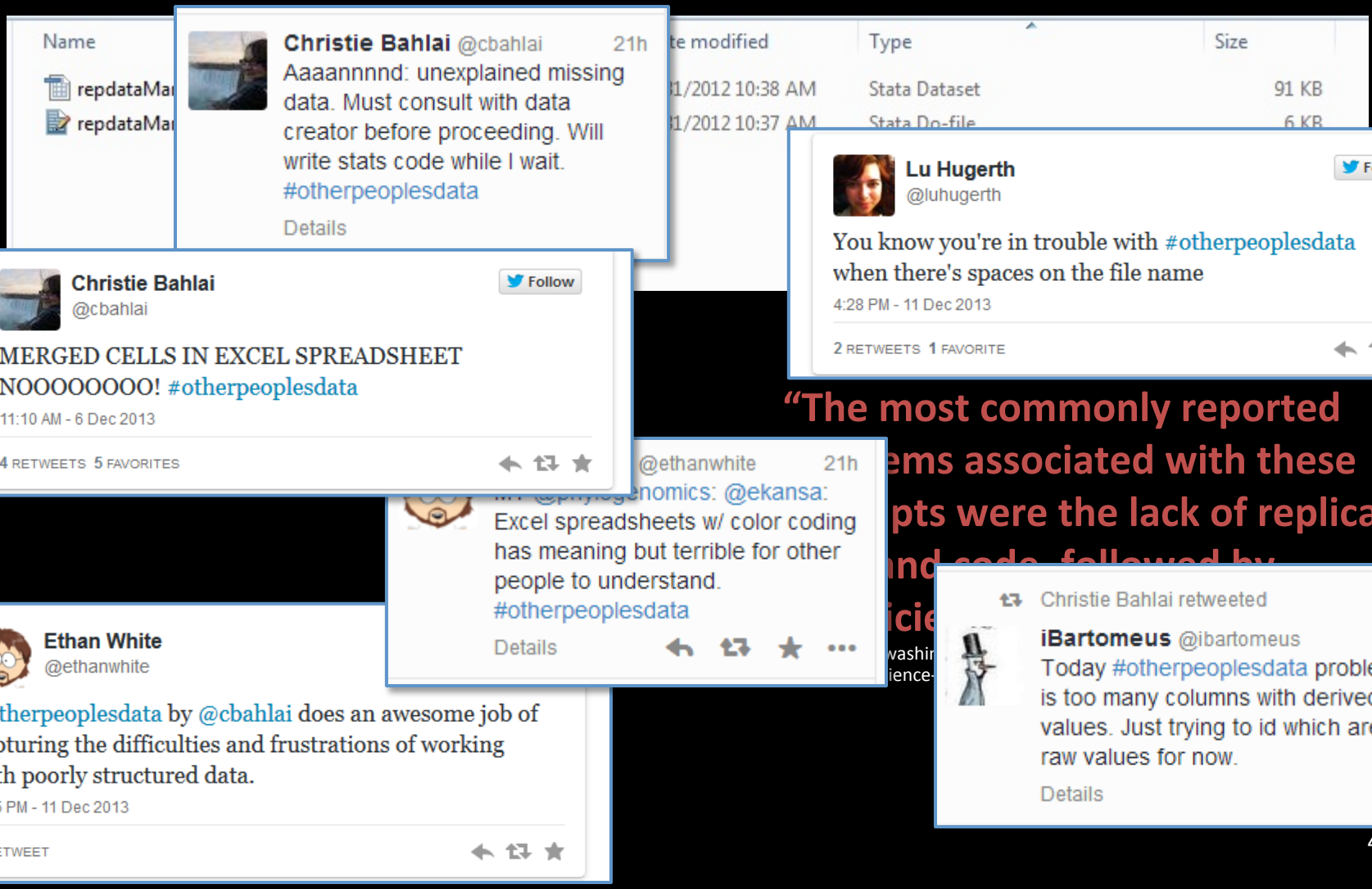
outline

- Data Quality
- Building the DDI
- Shifts
- Crisis of Quality & Loss of Data
- Committing to Data Quality

Data Sharing

Source: Nature (GARY WATERS/IKON IMAGES/CORBIS)

Challenges of Data Sharing



Intelligent Openness

Openness in itself has no value unless it is "intelligent openness."

This means that published data should be

- accessible (can they be readily located?),
- intelligible (can they be understood?),
- assessable (can their source and reliability be evaluated?) and
- The data must be usable by others (do the data have all the associated information required for reuse?).

Royal Society. Sceince Policy Centre. Science as an Open Enterprise. 2012.

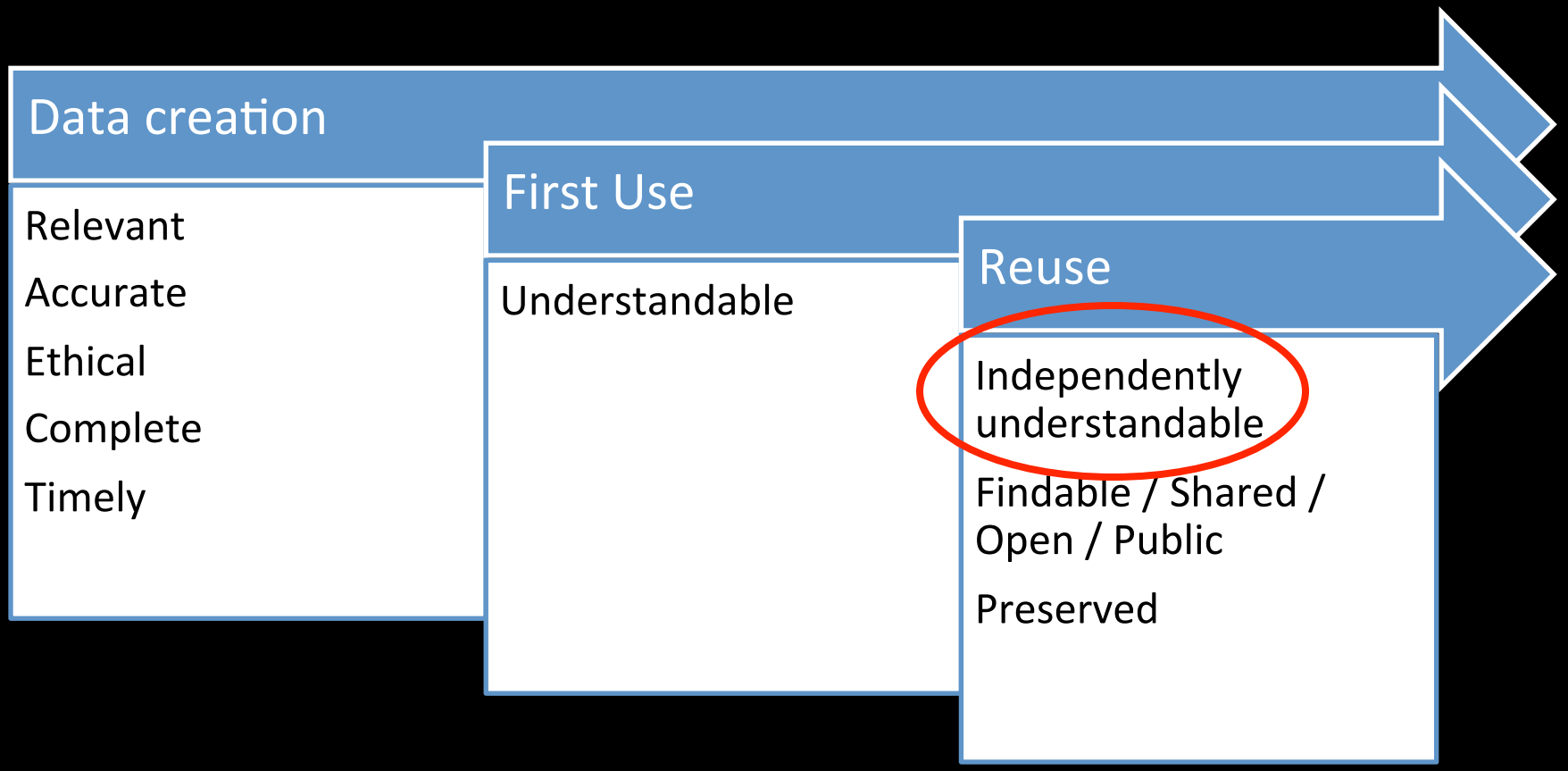
Independently
Understandable & Usable

http://yourprojectmanager.com.au/wp-content/uploads/2013/10/Focus-on-Quality.jpg

Other Measures of Quality

- assessable
- accessible
- quality of research
- quality of analysis or interpretation
- quality of instruments

Aspects of Data Quality



CCSDS RECOMMENDED PRACTICE FOR AN OAIS REFERENCE MODEL

Global Community: An extended Consumer community, in the context of Federated Archives, that accesses the holdings of several Archives via one or more common Finding Aids.

Independently Understandable: A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

Information: Any type of ~~knowledge that can be exchanged~~. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the Representation Information).

Information Object: A Data Object together with its Representation Information.

Information Package: A logical container composed of optional Content Information and

Independently Understandable

"A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals."

Source: OAIS. 2012. p. 1-12

Data Science Journal, Volume 12, 13 September 2013

OUT OF CITE, OUT OF MIND:

THE CURRENT STATE OF PRACTICE, POLICY, AND TECHNOLOGY FOR THE CITATION OF DATA

CODATA-ICSTI Task Group on Data Citation Standards and Practices

Edited by Yvonne M. Socha

CODATA Citation Principles

4. Access: Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.

One of the “First Principles” listed in CODATA Out of Cite, Out of Mind.

Executive Office of the President

Ensure that… “the direct results of federally funded scientific research are made available to and useful for the public, industry, and scientific community.”

Office of Science and Technology Policy

Memo to the Heads of Executive Departments and Agencies. Feb 2013

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus

OSTP definition of Data

Data = the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications

DATA NECESSARY FOR VALIDATION OF RESEARCH FINDINGS

What does USABLE mean?

- How usable does data need to be?
- How understandable does it need to be?
- What is required to make data reusable?
  - Expertise?
  - Previous knowledge of related data?
  - Need to go to other sources for context and further information?

DDI and usability

There is a history of documentation that provides what is required for data to be usable and understandable.

And the DDI was built upon that history.

Time machine to 1995

BUILDING THE DDI

Elements required for

using and understanding data.

Vardigan and Miller

"Within the social science community there was a recognized need for high-quality documentation so that **secondary investigators could understand and use the data…**"

codebooks

- Code / coding
- Books: Printed compilation
- Distributed as a publication with citation, table of contents, appendices

The NORC General Social Survey

A User's Guide

A National Survey of
FAMILIES and HOUSEHOLDS

INTRODUCTION

CODEBOOKS:

- Main Interview
- Self-Administered
- Husband/Wife (Partner)

台灣地區社會變遷基本調查計劃

第二期　第一、二次調查計劃執行報告

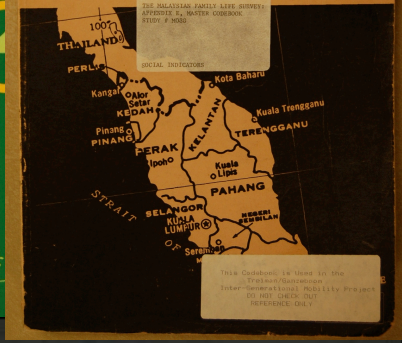
THE MALAYSIAN FAMILY LIFE SURVEY:
APPENDIX E, MASTER CODEBOOK

PREPARED FOR THE AGENCY FOR INTERNATIONAL DEVELOPMENT

TERRY FAIN, TAN POH KHEONG

R-2351/5-AID

JANUARY 1982



Public Use Data Tape Documentation

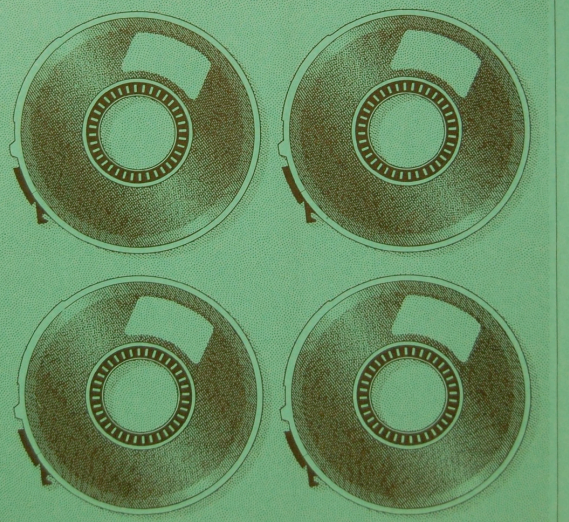
1986 Fetal Deaths Detail Record

National Center for Health Statistics

VITAL STATISTICS MORTALITY: FETAL DEATHS DETAIL RECORD, 1986
STUDY # H1900

HEALTH STATISTICS

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES • Public Health Service • Centers for Disease Control • National Center for Health Statistics



INTER-UNIVERSITY CONSORTIUM FOR POLITICAL RESEA

ICPR

The SRC 1960
American National
Election Study

September-December 1960

PRINCIPAL INVESTIGATORS: Angus Campbell, Philip Converse, Warren Miller, Donald Stokes

Institute for Social Research
University of Michigan

REVISED ICPR EDITION
SECOND PRINTING, 1974

SRA
ICPR
#7216

The Panel Study of Income Dynamics

A User's Guide

Guide to Major Social Science Data Bases 2

Martha S. Hill

building the DDI

Intellectual components from:

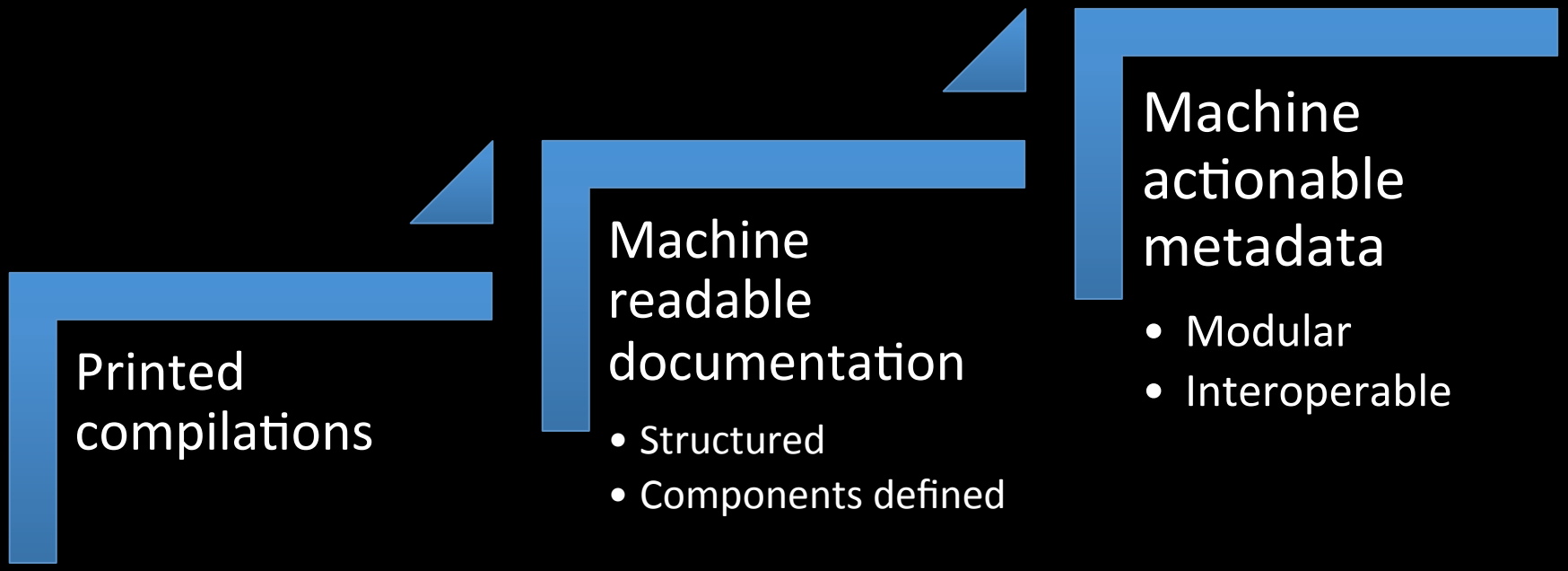
- Codebooks
- Study descriptions from data archives
- Bibliographic records establishing identity
- Statistical analysis metadata

Quality Documentation

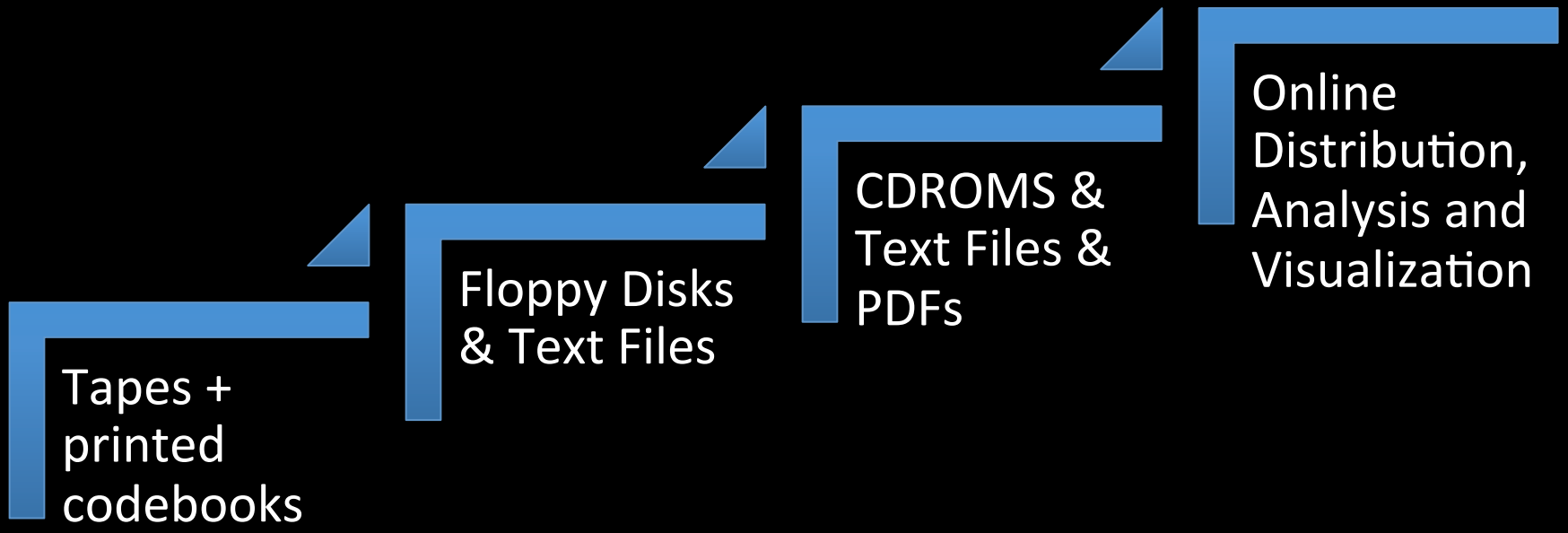
- We knew what good documentation was
- We consolidated what was good into a set of elements for the DDI and wrote a Tag Library
- We put it all into SGML then XML and the DDI Alliance was formed

But in the meantime there were major and significant SHIFTS to the social science research data universe

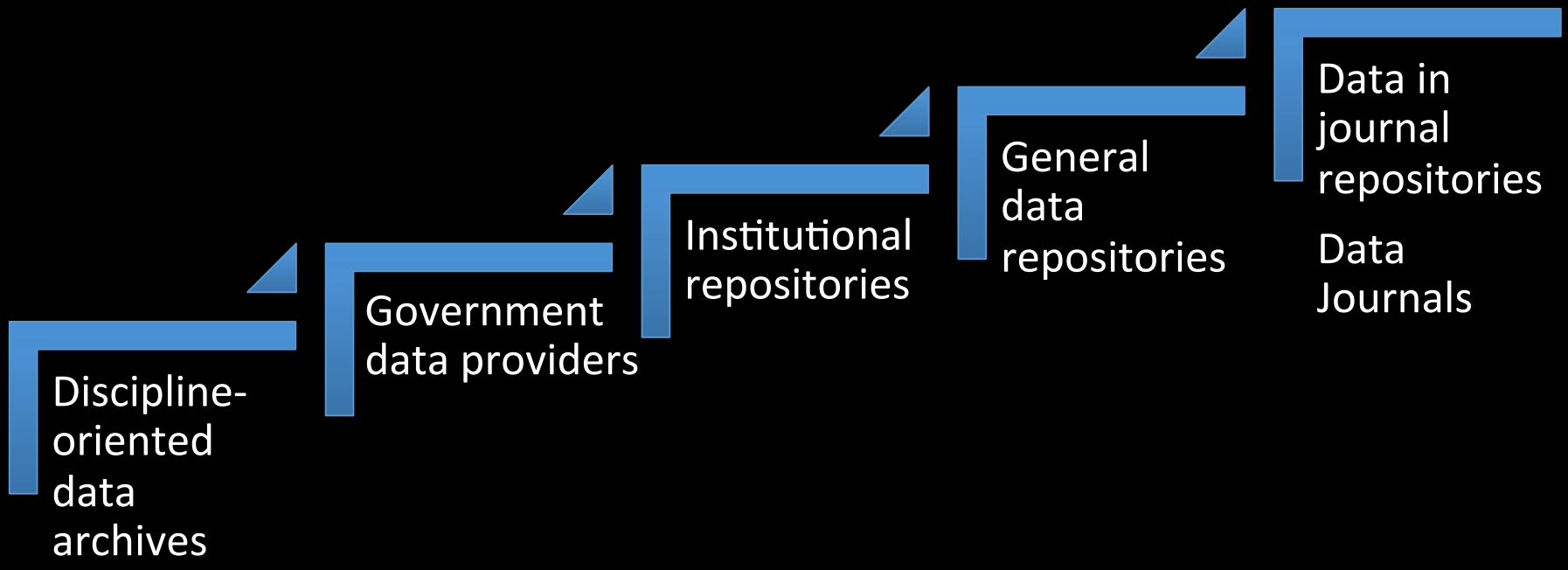
SHIFTS in DOCUMENTATION



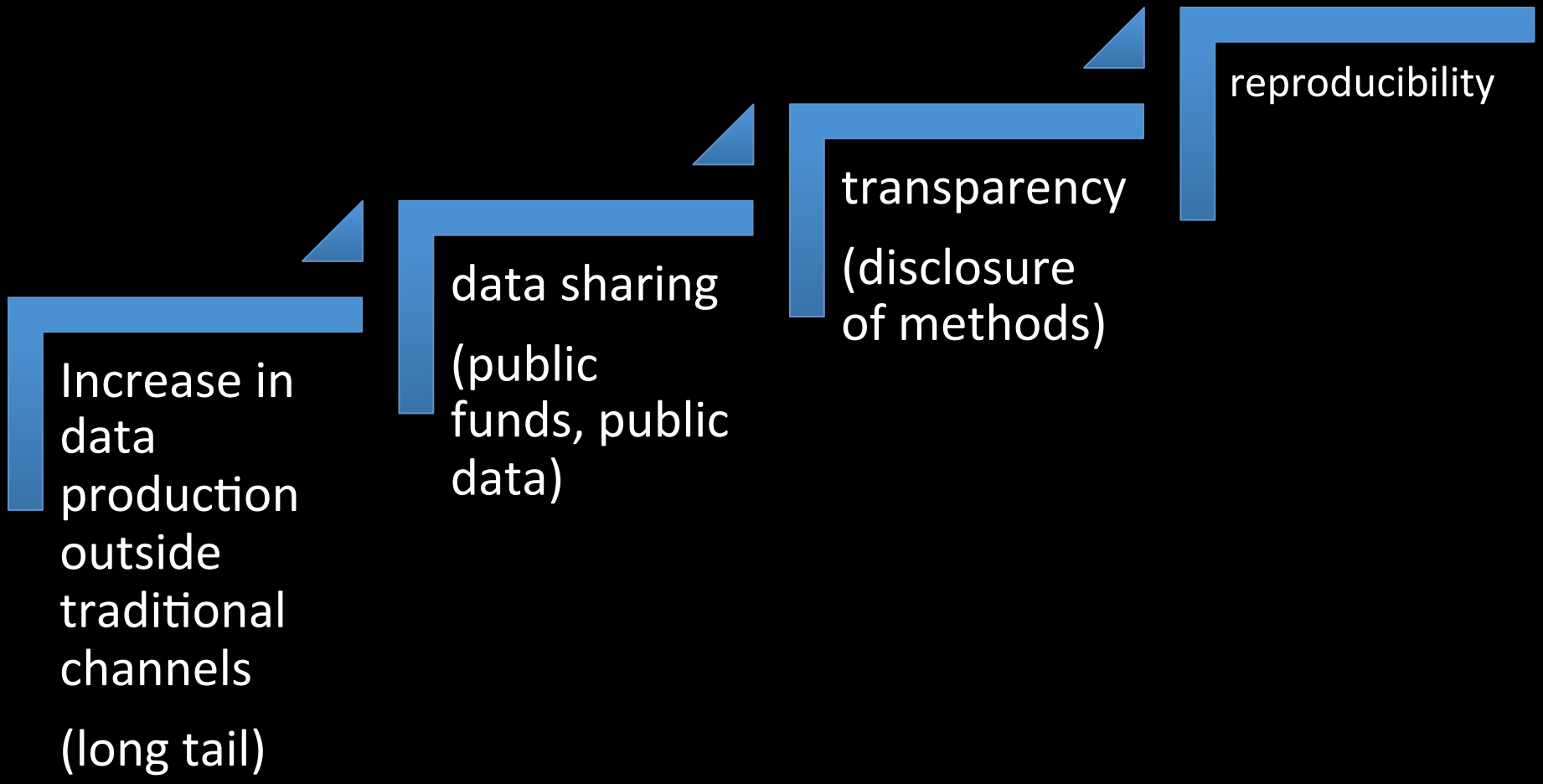
SHIFTS in STORAGE and ACCESS



SHIFTS in DATA ECOSYSTEM



SHIFTS in RESEARCH



SCIENCE FRICTION

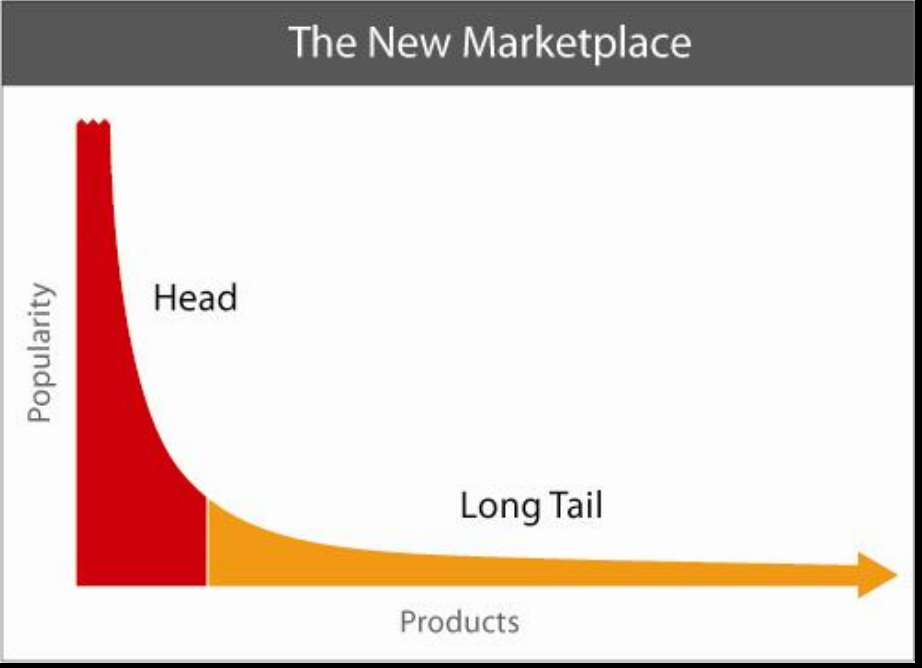
Between

the Pressure to share

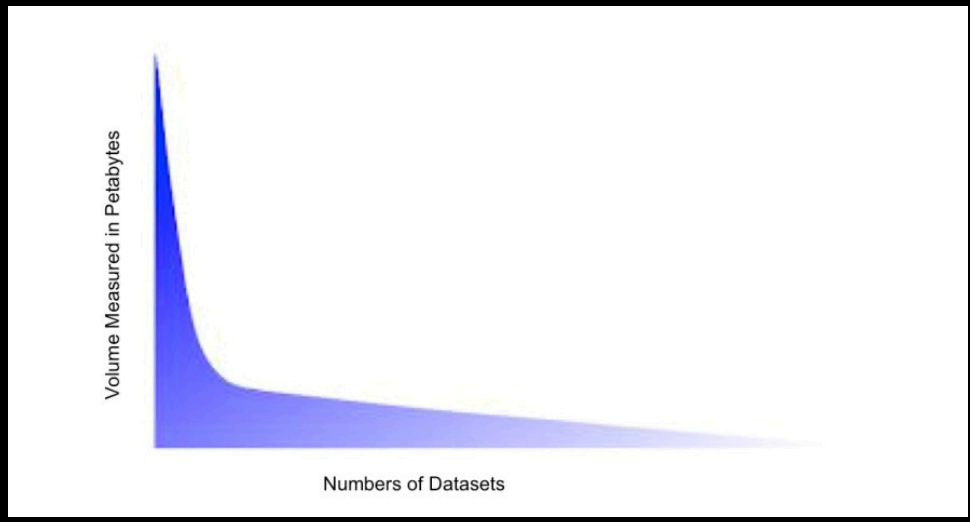
and

the Demands of sharing

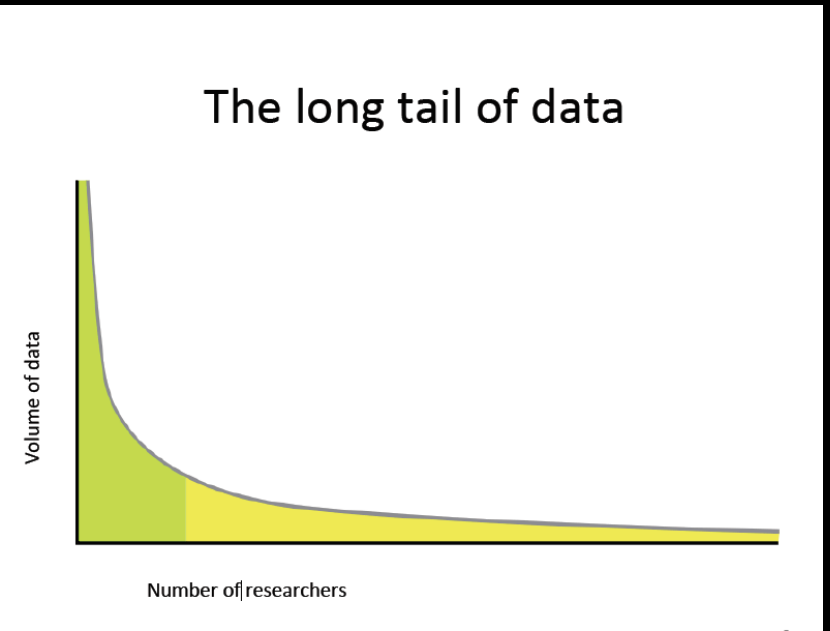
Long Tail of Data



http://www.thelongtail.com/conceptual.jpg



http://preservingresearchdataincanada.net/2012/12/05/hello-world/longtaildata1/



http://works.bepress.com/borgman/269/

Typical characteristics of 'long tail' science data

- Excel spreadsheets
- File names as identifiers
- Unlabeled rows and columns
- Code and data files not managed under version control
- Sometimes data are pushed to large domain databases or archives but usually stay 'in the lab'
- Shared informally or without curation

conundrum

Researcher resistance

- Concerns about additional work and time
- An acceptable workflow needs to be created
- Lack of awareness about metadata standards for data publication
- Unclear how to put routine best-practice data management in place

Source: Out of Cite Out of Mind. 6.3.6

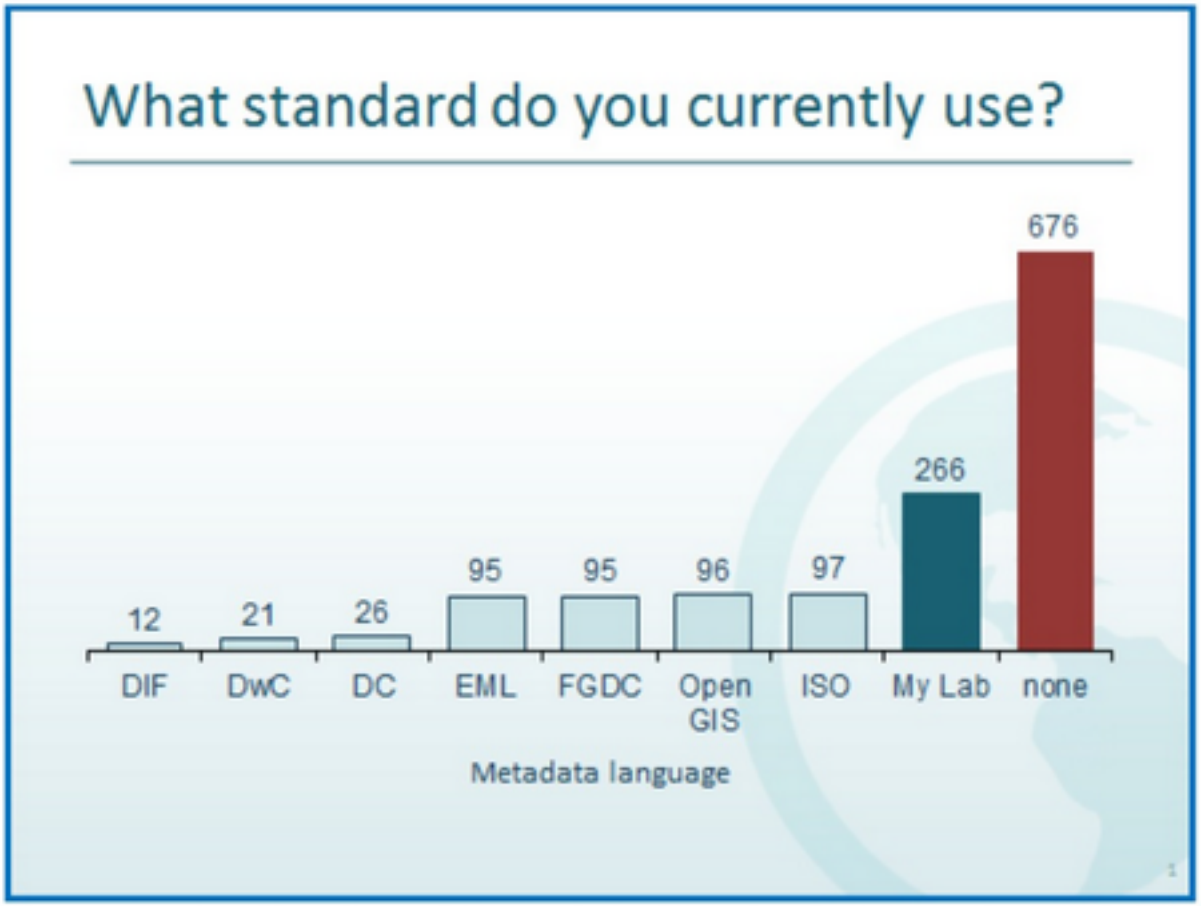


Figure 1: Application of Metadata Standards (Tenopir, et al, 2011), *n* = 1329.

Documentation Problems

- First problem is that documentation isn't being created at all -- not at the point of data creation, during processing, or after a project has finished.
- Second problem is that documentation is of poor quality (the readme file problem).
- Third problem is documentation that does exist isn't being compiled (or linked). Intellectual components aren't gathered together. No more 'codebooks.'

DATA AT RISK

CAN'T USE IT
CAN'T UNDERSTAND IT

- We understand quite a bit about digital preservation environments and what it takes to make a repository TRUSTWORTHY.

- But we know much less about how to make the digital objects in those environments meet the test of time in terms of USABILITY

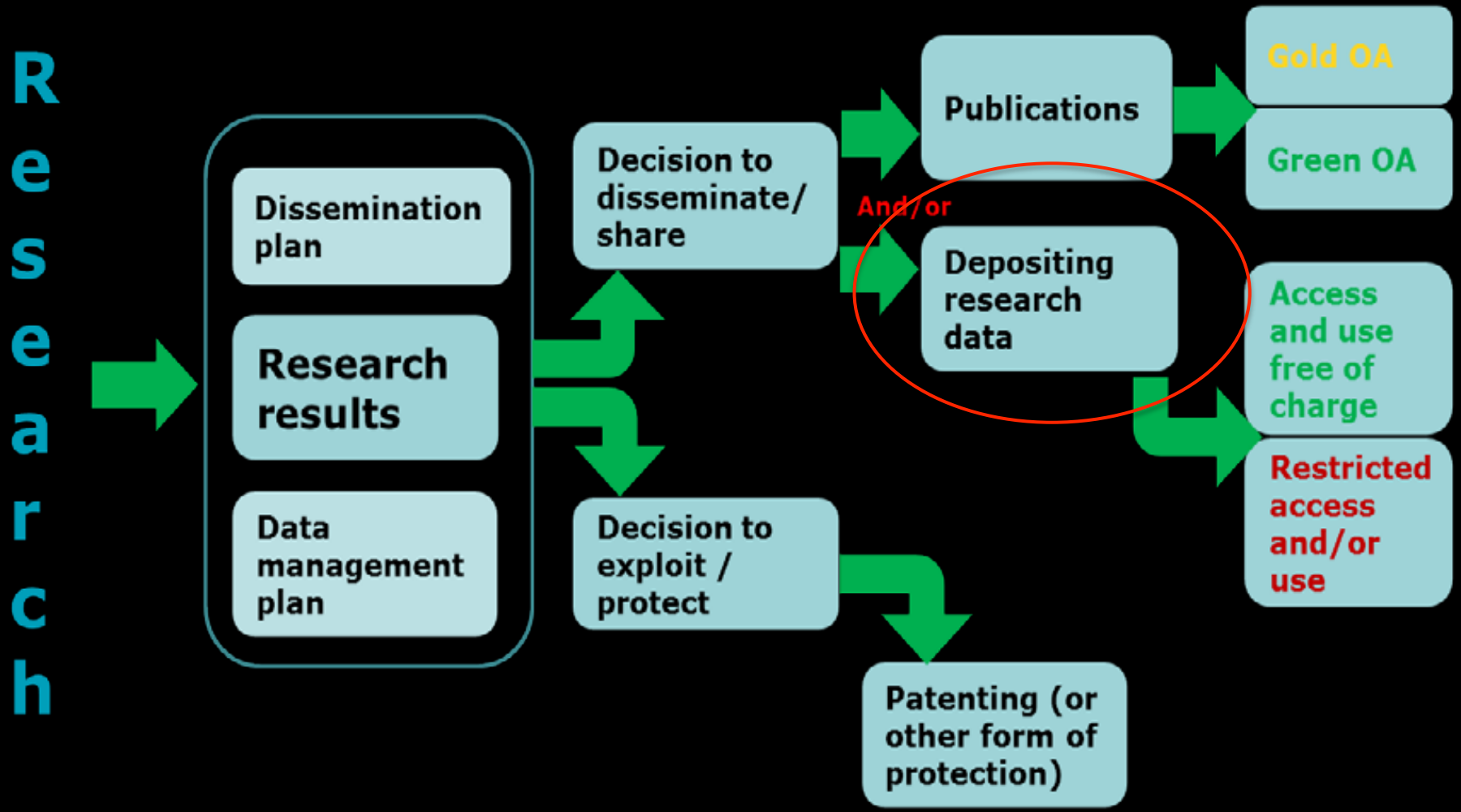
Unusable data = lost data

Image: Shutterstock.com/Lightspring
From:
http://slashdot.org/topic/datacenter/neglect-causes-massive-loss-of-irreplaceable-research-data/

What would help improve data quality?

1. Data Quality Review during the Curatorial Process
2. Data Quality during the Research Process

DATA QUALITY REVIEW



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation ***Source:*** European Commission. The European Framework Programme for Research and Innovation. Horizon 2020 Guidelines on Open Access to Scientific Publications and Research Data. 2013

Who reviews data quality at the Deposit Stage?

DATA ARCHIVES

Intentionally process data for reuse

Social Science Data Archive

UNC THE ODUM INSTITUTE Data Archive

cessda

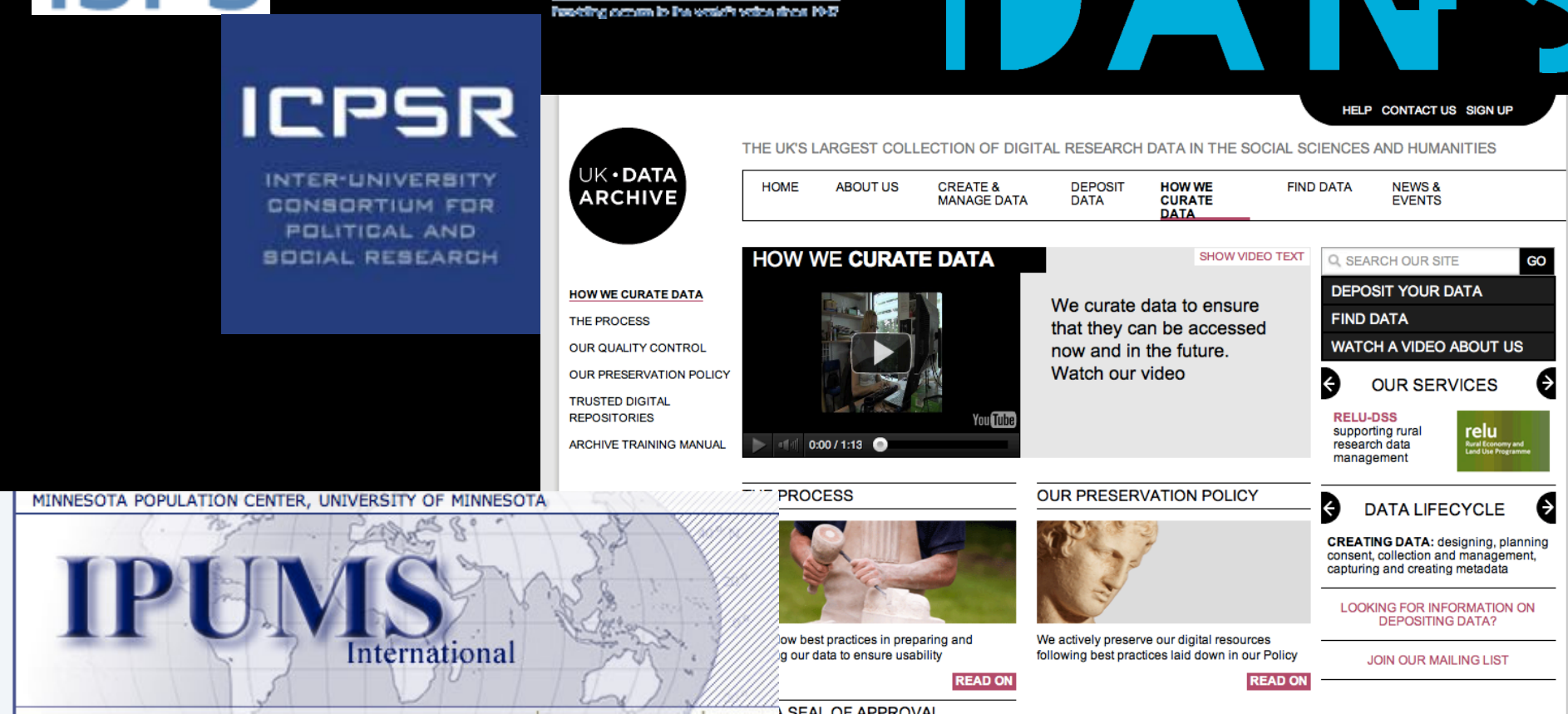
gesis

Yale ISPS

ROPER CENTER FOR PUBLIC OPINION ARCHIVES

DANS

ICPSR INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH



MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA

IPUMS International

Data quality review process

REVIEW FILES  Assign persistent IDs * Create a citation to the study and a study level metadata record * Record file details (size, format, checksums) * Check that all files are present * Verify that content of files matches expected format * Create non-proprietary versions of the files * Implement migration strategy for file formats * Monitor bits	**REVIEW DATA**  Check for undocumented variable and value information or out of range codes * Review data for confidentiality issues
REVIEW DOCUMENTATION  Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products	**REVIEW CODE**  Check and verify code for data analysis and replication

Source: Peer, Green, and Stephenson. 2014. Committing to Data Quality Review. International Journal of Digital Curation. Forthcoming.

Preprint: http://isps.yale.edu/sites/default/files/files/CommitingToDataQualityReview_idcc14-PrePrint.pdf

New service with 2 options

- Professional Curation: storage, persistent ID, plus ICPSR curators will review data documentation and formats and make it Open
- Self deposit: $600 includes 10 years of storage, persistent ID, catalog type metadata (no format migration, no checking data or documentation) No data quality review.

INCREASE in DIY

- Do it yourself data collection and documentation
- Do it yourself data management
- Do it yourself data “archiving” and “preservation”

DIY and Repositories

- Institutional repositories do not take on the role of data quality review, as defined here.
- Dataverse: tools to support DIY review of content
- Dryad: curatorial verification of formats and study level metadata
- Figshare: no review (by definition of service)

DQR process in other repositories

REVIEW FILES
Create persistent ID
Record file sizes and formats
Create checksums
Check for completeness, confirm all files are present (data, and required documentation and code if available)
Create study-level metadata record including file information
Create citation
Create non-proprietary file formats for preservation
REVIEW DOCUMENTATION
Confirm comprehensive descriptive information
Confirm methodology and sampling information
Create documentation compliant with community standards, e.g., DDI XML
REVIEW DATA
Run frequencies and check for undocumented or out of range codes
Standardize missing values; check for consistency and skip patterns
Check and edit variable and value labels
Check and add question wording (surveys)
Review data for confidentiality issues; Recode variables to address confidentiality concerns
Generate multiple data formats for dissemination
REVIEW CODE
Check and verify replication code
PUBLISH & LINK
Publish to access system
Link to other research products (e.g., publications, registries, grants)
PRESERVE
Migration strategy for file formats
Monitor bits

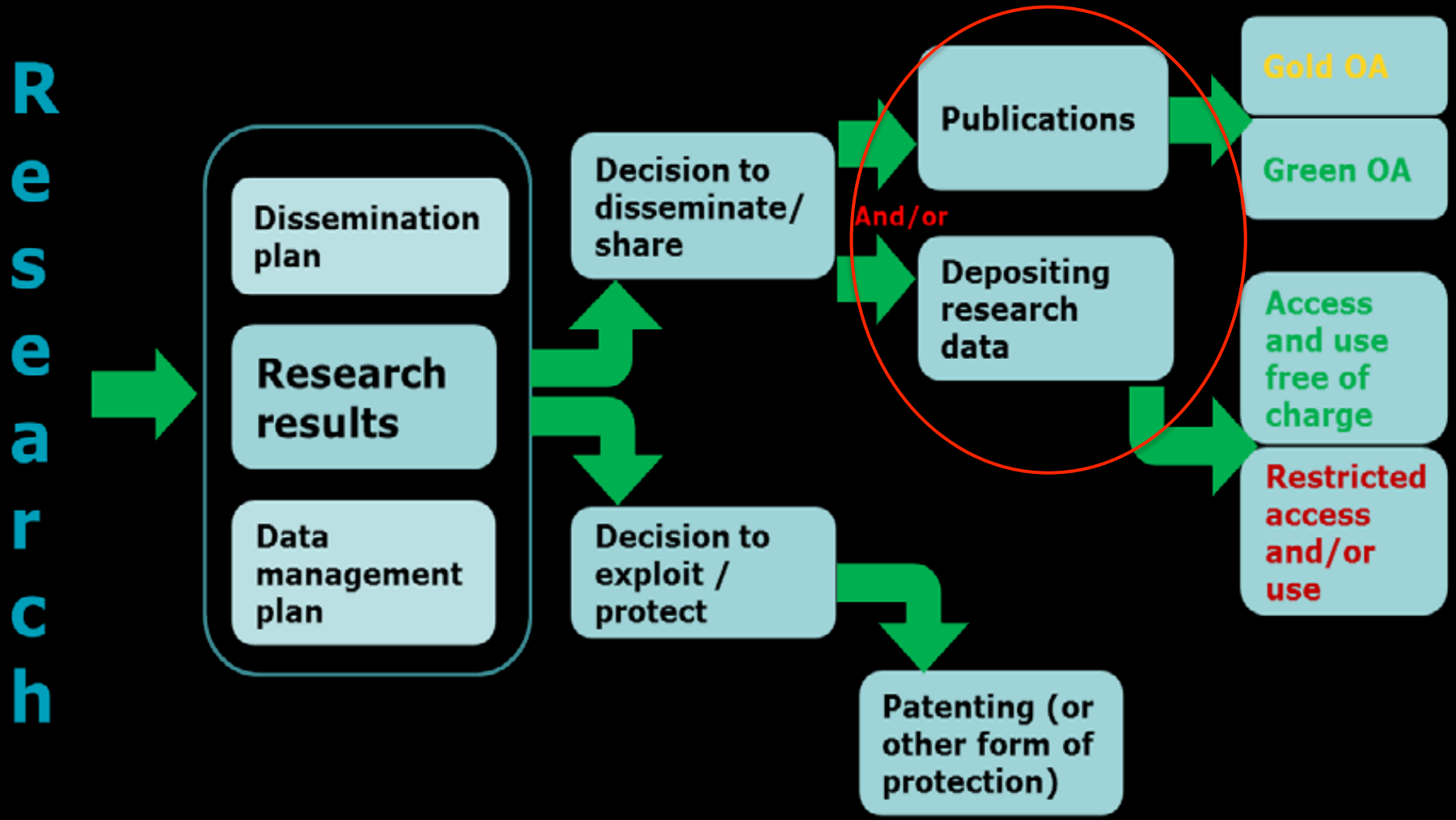
Expert • Independent • Nonprofit

ConsumerReports.org®

Testing products since 1936.

Consumer Reports is an **expert, independent, nonprofit organization** whose mission is to work for a fair, just, and safe marketplace for all consumers and to empower consumers to protect themselves.

Report a Safety Problem | Donate



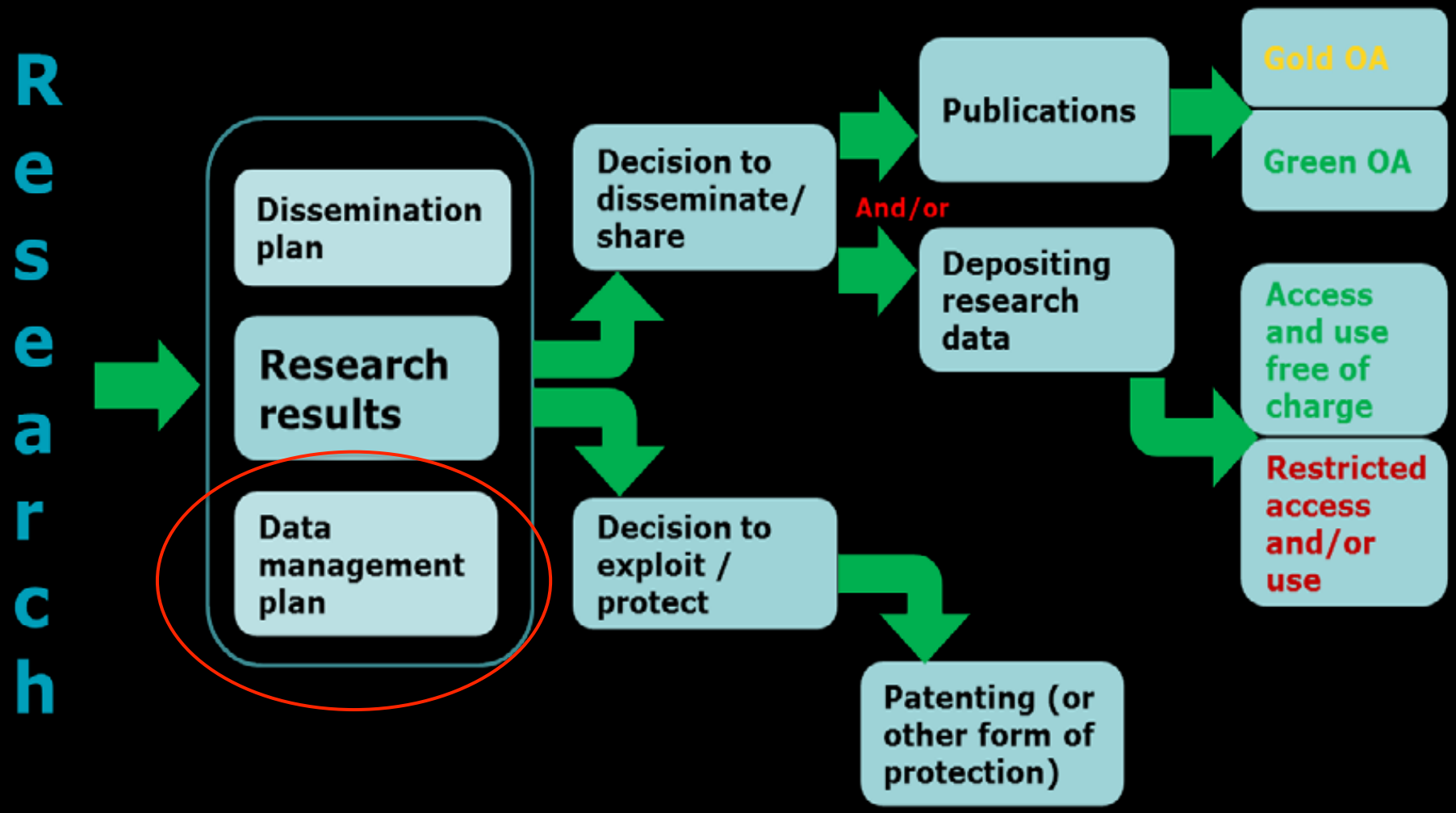
Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

Data Quality Review and Scholarly Journals

- Data quality peer review
- Data descriptors
- Data papers
- Data journals
- Dryad model – partner with journals

See also Marieke Guy & Monica Duke (DCC) at IASSIST 2013 on the rise of data journals

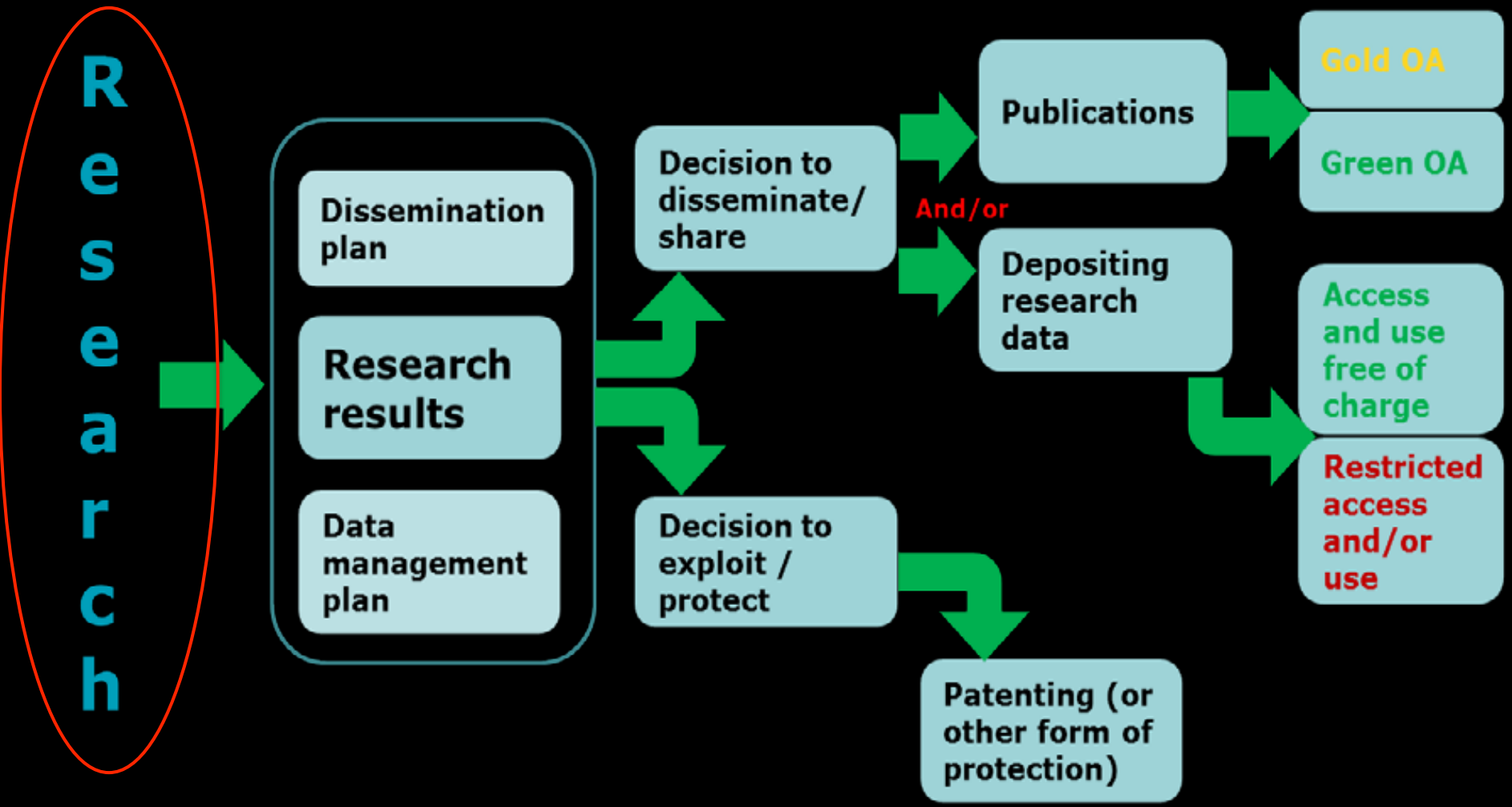
Data Quality and the Research Process



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

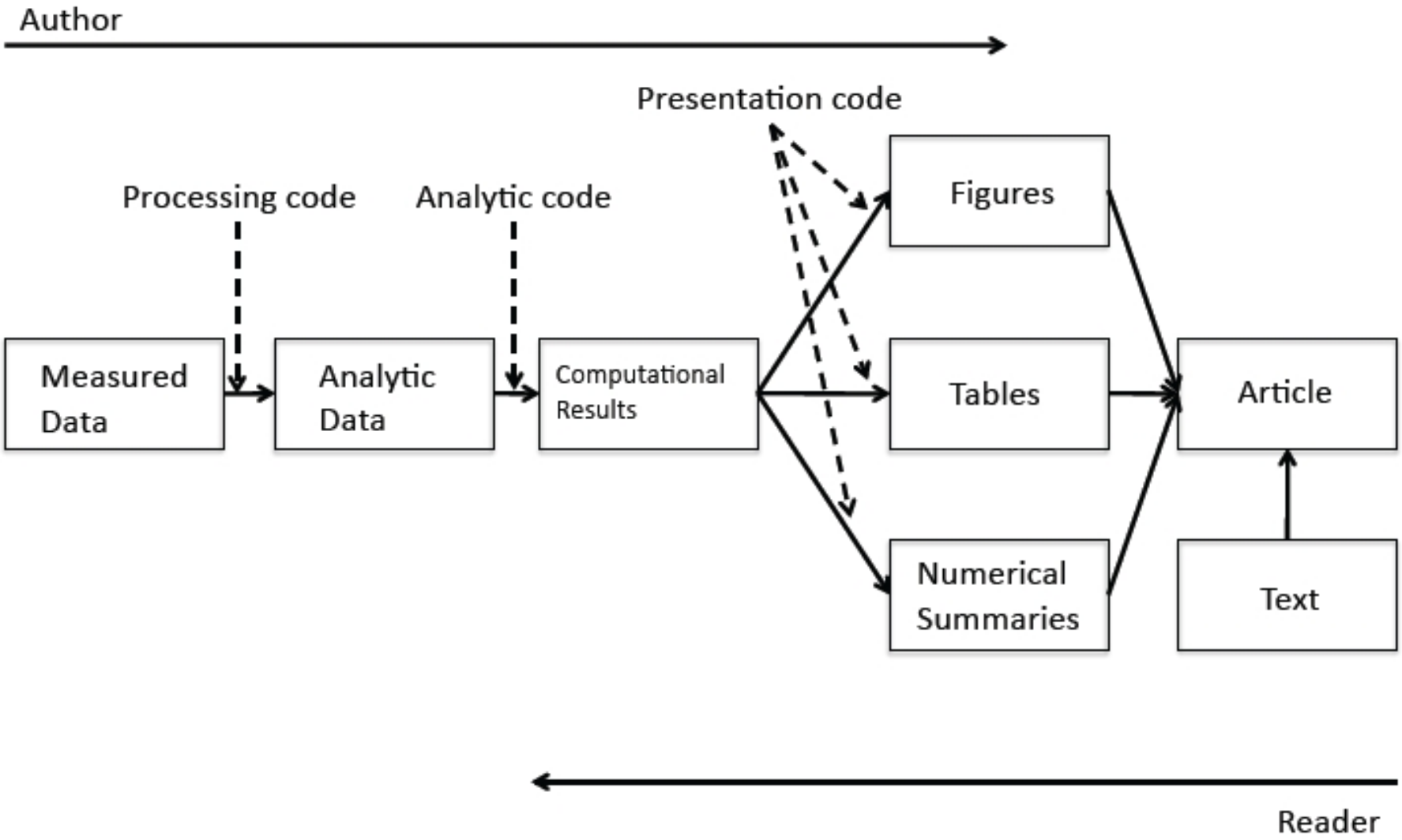
Data management plans

- Address downstream data production for open sharing of funded research.
- Great step forward.
- BUT do DMPs cover REVIEWING measures of data quality in regard to understandability?
  - NOT REALLY
- Is there ANYTHING in a DMP about who will review the quality of data?
  - NOT REALLY



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

Expose more of the research workflow



http://www.stodden.net/AMP2011/slides/pengslides.pdf

Sources of “Understandability” for informed reuse

Research artifacts describe the data, code, analysis

- Informal documentation
- Workflow tools
- Data papers
- References and citations; publications
- Annotations by secondary users

Project Hosting

- GitHub
- Bitbucket
- Dat-data.com

Social Science example:

Open Science Framework

Open Science Framework

Public Watch 5 15

Contributors: Joshua Carp | Jeffrey R. Spies | Nan Chen | Steve Loria | Lyndsy Simon | Chris Seto | Johanna Cohoon | Denise Holman | Tanesha Hudson | Melissa Lewis | Sam Portnow | Jacob Rosenberg | Harry Rybacki | Andrew Sallans | Alex Schiller | Brian A. Nosek

Date Created: 5/31/2012 1:36 AM | Last Updated: 3/10/2014 10:49 AM

The Open Science Framework (OSF) is part network of research materials, part version control system, and part collaboration software. The purpose of the software is to support the scientist's workflow and help increase the alignment between scientific values and scientific practices.

1. **Document and archive studies.** Move the organization and management of study materials from the desktop into the cloud. Labs can organize, share, and archive study materials among team members. Web-based project management reduces the likelihood of losing study materials due to computer malfunction, changing personnel, or just forgetting where you put the damn thing.
2. **Share and find materials.** With a click, make study materials public so that other researchers can find, use and cite them. Find materials by other researchers to avoid reinventing something that already exists.
3. **Detail individual contribution.** Assign citable, contributor credit to any research material - tools, analysis scripts, methods, measures, data.
4. **Increase transparency.** Make as much of the scientific workflow public as desired - as it is developed or after publication of reports. Find public projects here.
5. **Registration.** Registering materials can certify what was done in advance of data analysis, or confirm the exact state of the project at important points of the lifecycle such as manuscript submission or at the onset of data collection. Discover public registrations here.
6. **Manage scientific workflow.** A structured, flexible system can provide efficiency gain to workflow and clarity to project objectives, as pictured.

Name: home
Version: 16 (current)

Edit History New Page

Project Wiki Pages

home

open_science_projects

Component Wiki Pages

Developer API (other)

Demo Add-Ons (other)

Mirroring (other)

Sync GitHub to figshare

- Using version control that figshare provides, allows researchers to cite the exact files that were used in generating research outputs.
- A copy of all files is also pulled from GitHub into figshare to ensure the persistence of the code as a research output.

See more at: http://figshare.com/blog/Upload_your_code_and_thesis_to_figshare_/118#sthash.aRlHXGbo.dpuf

R to Dataverse

- Move reproducible research reports to Dataverse directly from R
- dvn wrappers for the Data Sharing search utility and Data Deposit (built on SWORD protocol)
- Dataverse offers Version control, DOIs, DDI and DC metadata
- Process: Create study, build metadata, add files, release to public

DDI and Data Quality

- Documentation = Best of the best
- Lifecycle model
- Tools to support best practices by researchers
- Tools to support data quality review by curators & publishers

- REUSE, REPRODUCE, REPLICATE

CHALLENGES

Critics say:

- DQR is too much work, no incentives or support
- Informal communications will suffice for usability
- Curation doesn't scale and is not sustainable

Need evidence of data loss and cost models for reviewing data

What will it take?

- Tools and Training
- Partnerships
- Policies
- Incentives
- Community
- Commitment to Independent and Informed usability over time

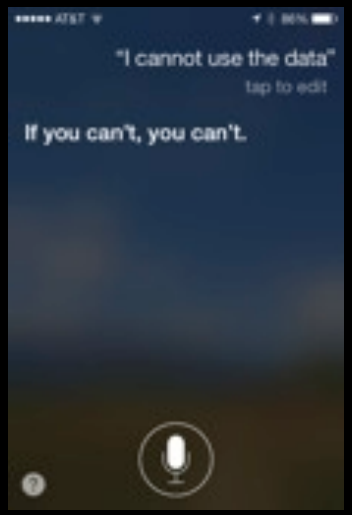
A community commitment

Improving the quality of data is an investment in future data sharing.

Improving the quality of the data is an obligation of *any* entity that assumes responsibility over the data.

It's in everyone's interest!

THANK YOU!



Special thanks to Limor Peer, PhD. Institution for Social and Policy Studies. Yale University.

Ann Green dlifecycle@gmail.com
Digital Lifecycle Research & Consulting
sites.google.com/site/dlifecycle/