Reproducible research at Census

Carl Lagoze (clagoze@umich.edu)

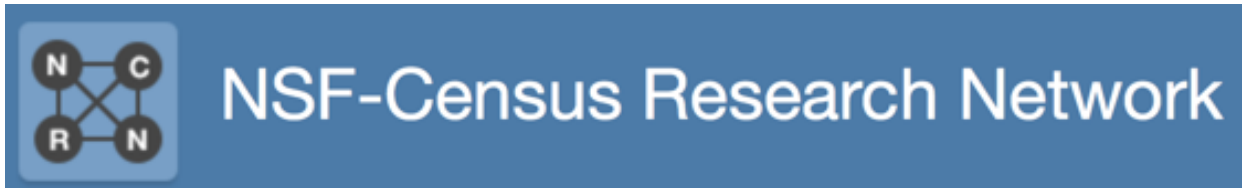
Associate Professor

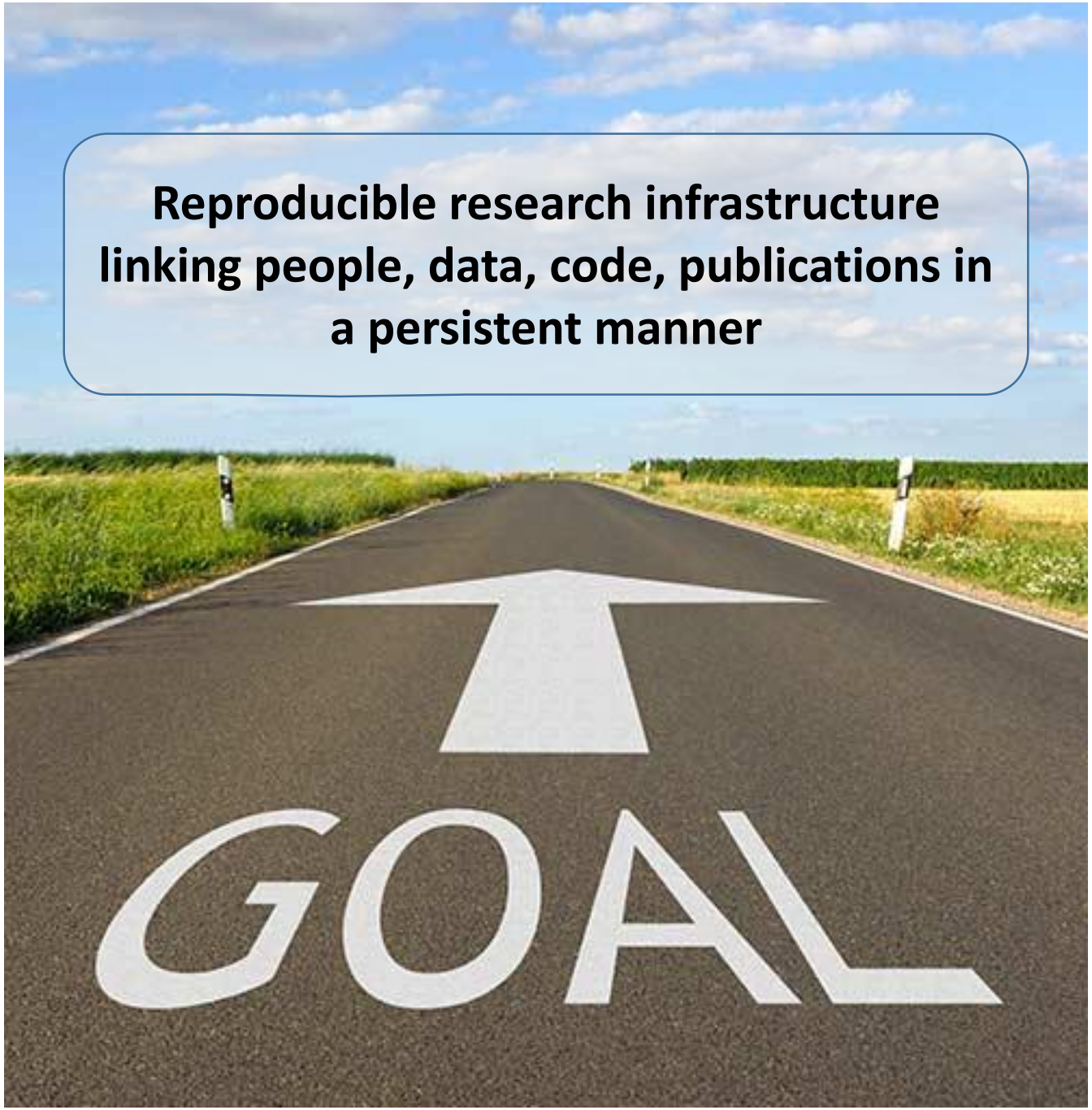
University of Michigan School of Information

Acknowledgements

Based on work with

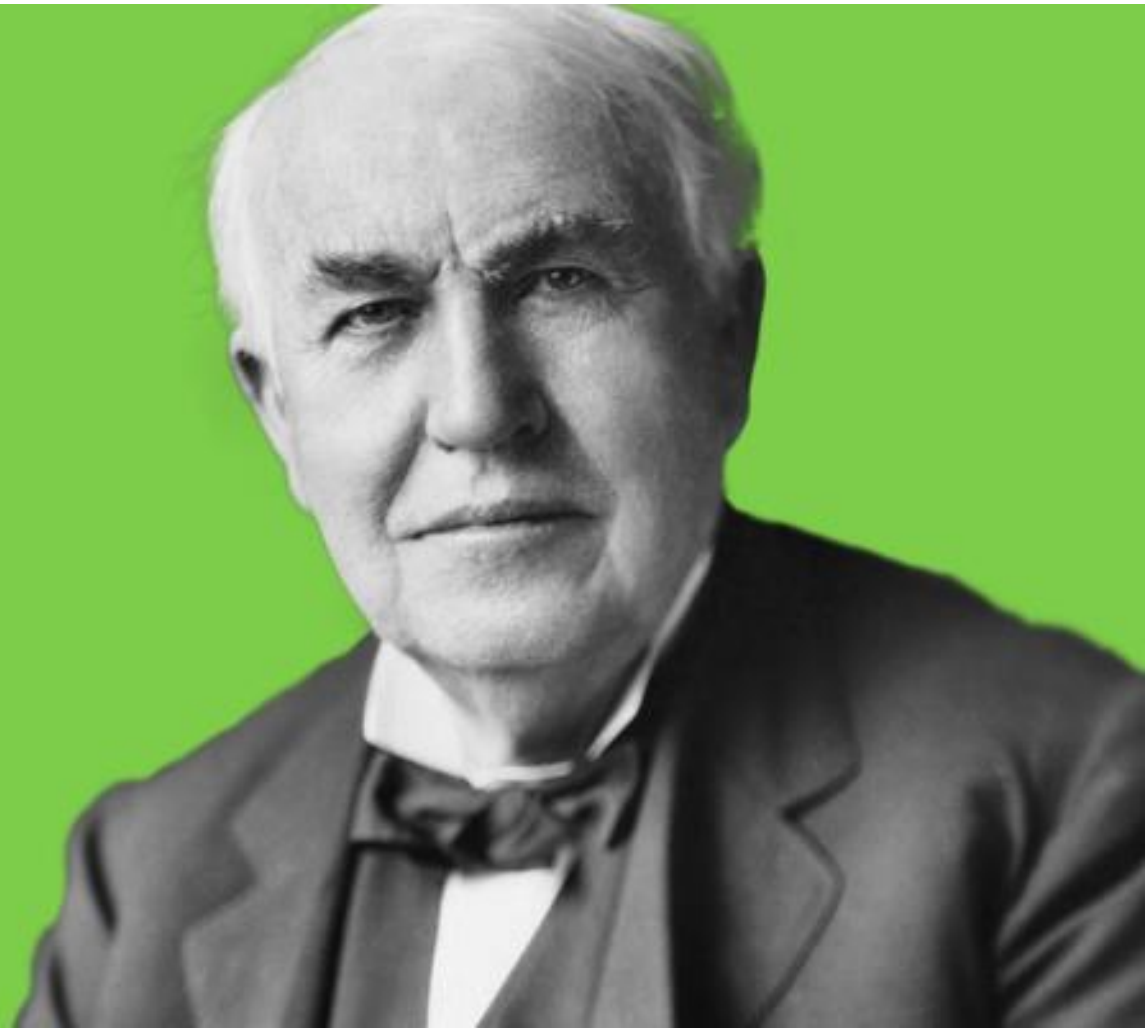
- Lars Vilhuber (Cornell University)
- William Block (Cornell University)
- Benjamin Perry (formerly Cornell University)
- Venkata Kambhampaty (formerly Cornell University)
- Kyle Brumsted (McGill University)
- Jeremy Williams (Cornell University)
- Carl Lagoze (University of Michigan)
- John Abowd (formerly Cornell University)




Reproducible research infrastructure linking people, data, code, publications in a persistent manner
GOAL

"I have not **FAILED.**
I've just found
10,000 WAYS
that won't work."

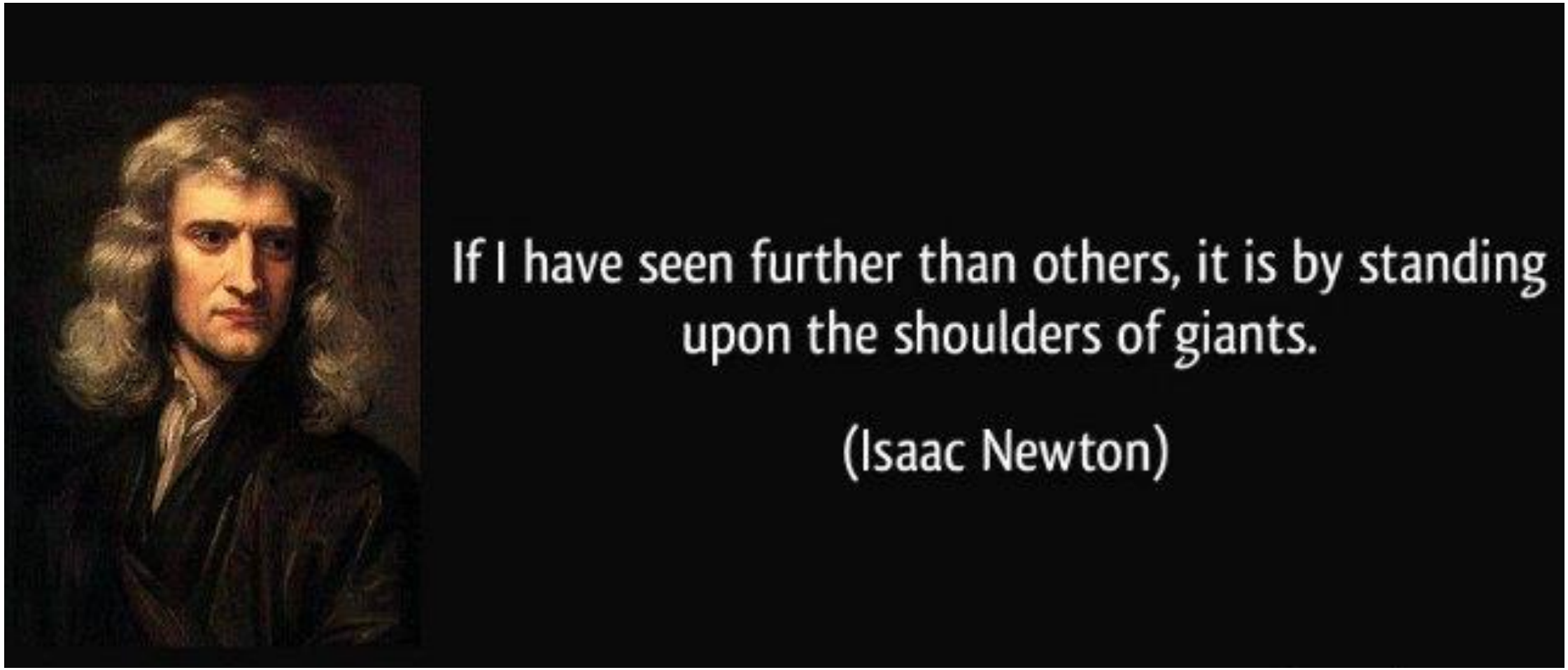
—Thomas Edison



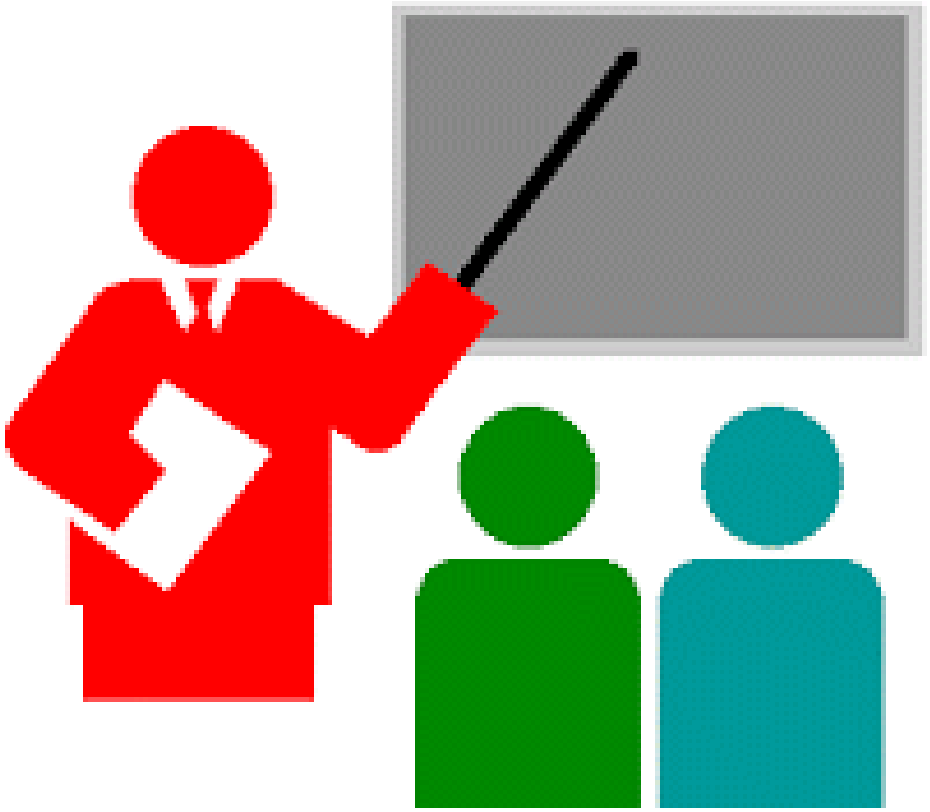
Why reproducibility?

Why reproducibility (1): Integrity of research

Why reproducibility (2): Advancement of Science



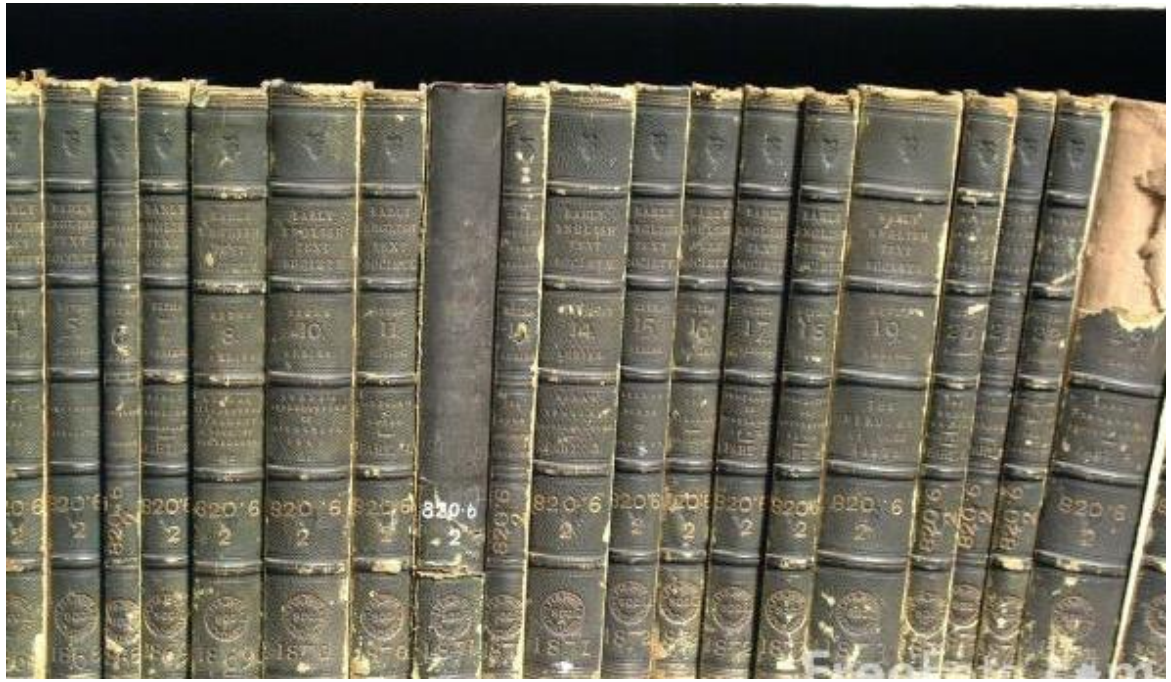
Why reproducibility (3): Education and Pedagogy



What’s getting in the way?

“… to date, online versions of ‘scholarly outputs’ have tended to replicate print forms, rather than exploit the additional functionalities afforded by the digital terrain. “

FORCE11 Manifesto - https://www.force11.org/about/manifesto



D-Lib Magazine
September 2004

Volume 10 Number 9

ISSN 1082-9873

Rethinking Scholarly Communication

Building the System that Scholars Deserve

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
<herbertv@lanl.gov>

Sandy Payette
Cornell University, Computing and Information Science
<payette@cs.cornell.edu>

John Erickson
Hewlett-Packard Laboratories, Digital Media Systems Lab
<john.erickson@hp.com>

Carl Lagoze
Cornell University, Computing and Information Science
<lagoze@cscornell.edu>

Simeon Warner
Cornell University, Computing and Information Science
<simeon@cs.cornell.edu>

“Our vision is based on our belief that the future scholarly communication system should closely resemble—and be intertwined with—the scholarly endeavor itself, rather than being its after-thought or annex. “

Vienna PRINCIPLES

a vision for scholarly communication

1 Accessibility
2 Discoverability
3 Reusability
4 Reproducibility
5 Transparency
6 Understandability
7 Collaboration
8 Quality Assurance
9 Evaluation
10 Validated Progress
11 Innovation
12 Public Good

https://zenodo.org/record/55597#.V3mbBldygl4

Back to the future?

Enthusiasm for the “revolution” varies

Reproducibility is a sociotechnical problem


andis

What are the barriers to progress?

Identification

Granularity

Transience

Data Sharing

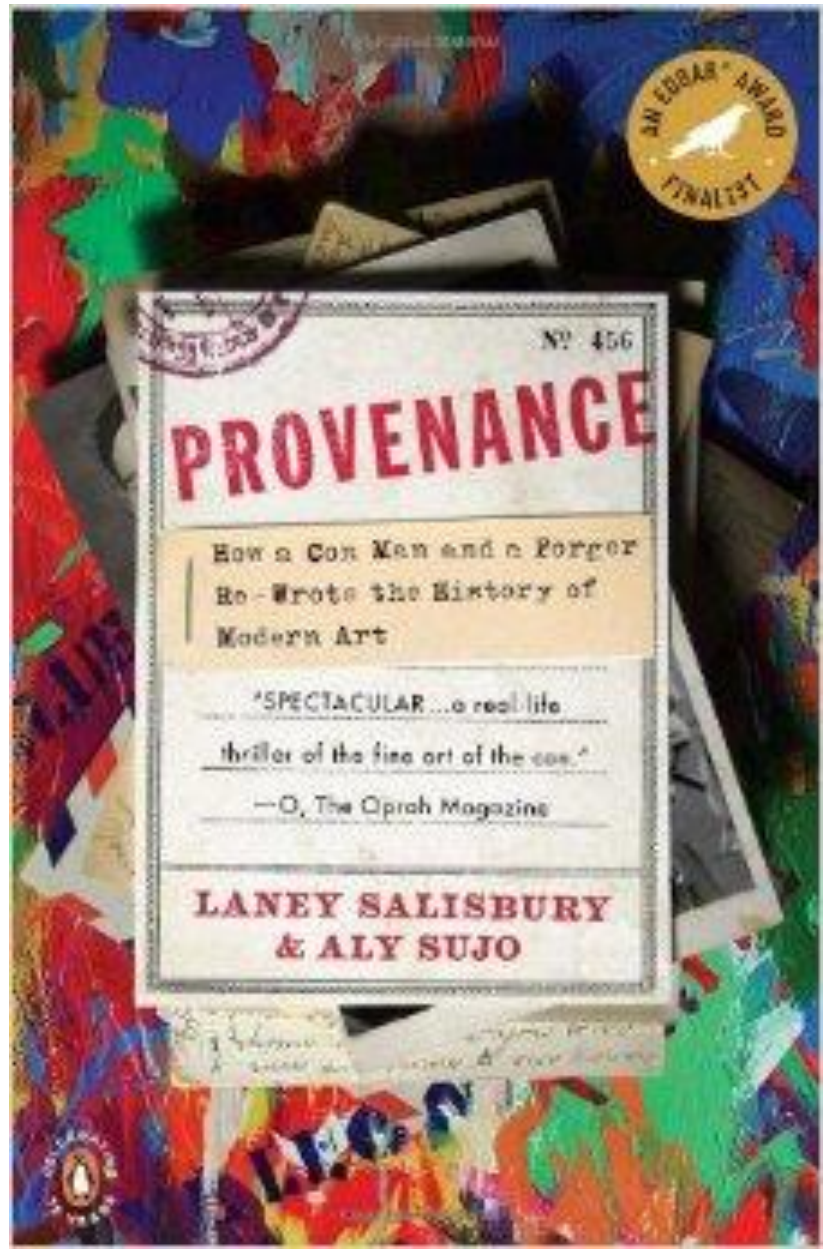
"Much more is understood about why researchers do not share data than about when, why, and how researchers do share data, or about when, how, and why researchers or the public reuse data."

Borgman, C. (2011). The Conundrum of Sharing Research Data. *Journal of the American Society for Information …*, 1–40. http://doi.org/10.2139/ssrn.1869155

Metadata


Privacy

Data integrity


AN EDGAR® AWARD FINALIST
№ 456
PROVENANCE
How a Con Man and a Forger Re-Wrote the History of Modern Art
"SPECTACULAR...a real-life thriller of the fine art of the con."
—O, The Oprah Magazine
LANEY SALISBURY & ALY SUJO

Low hanging fruit

Identity and Citation

Repositories

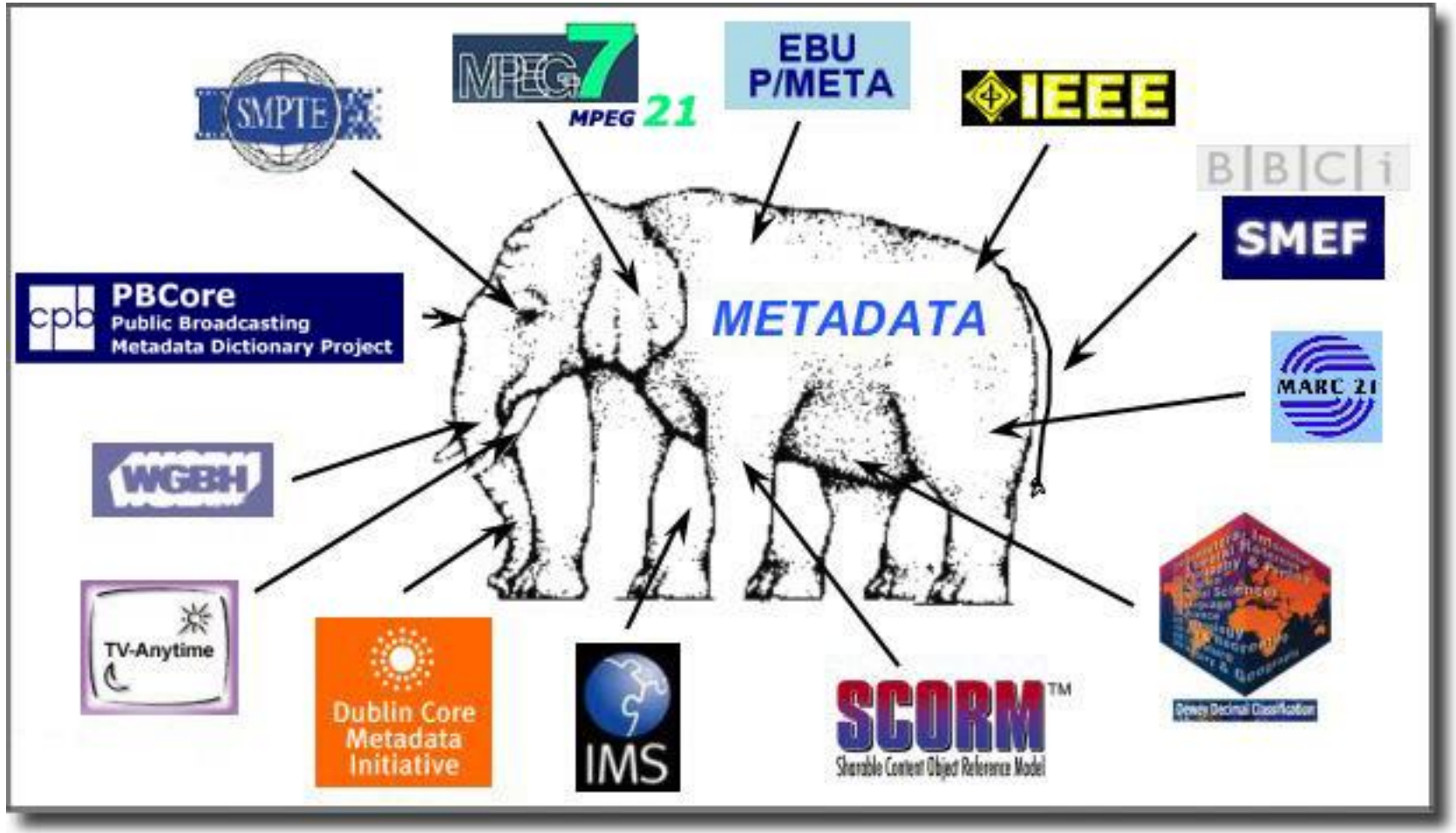
The Dataverse Project

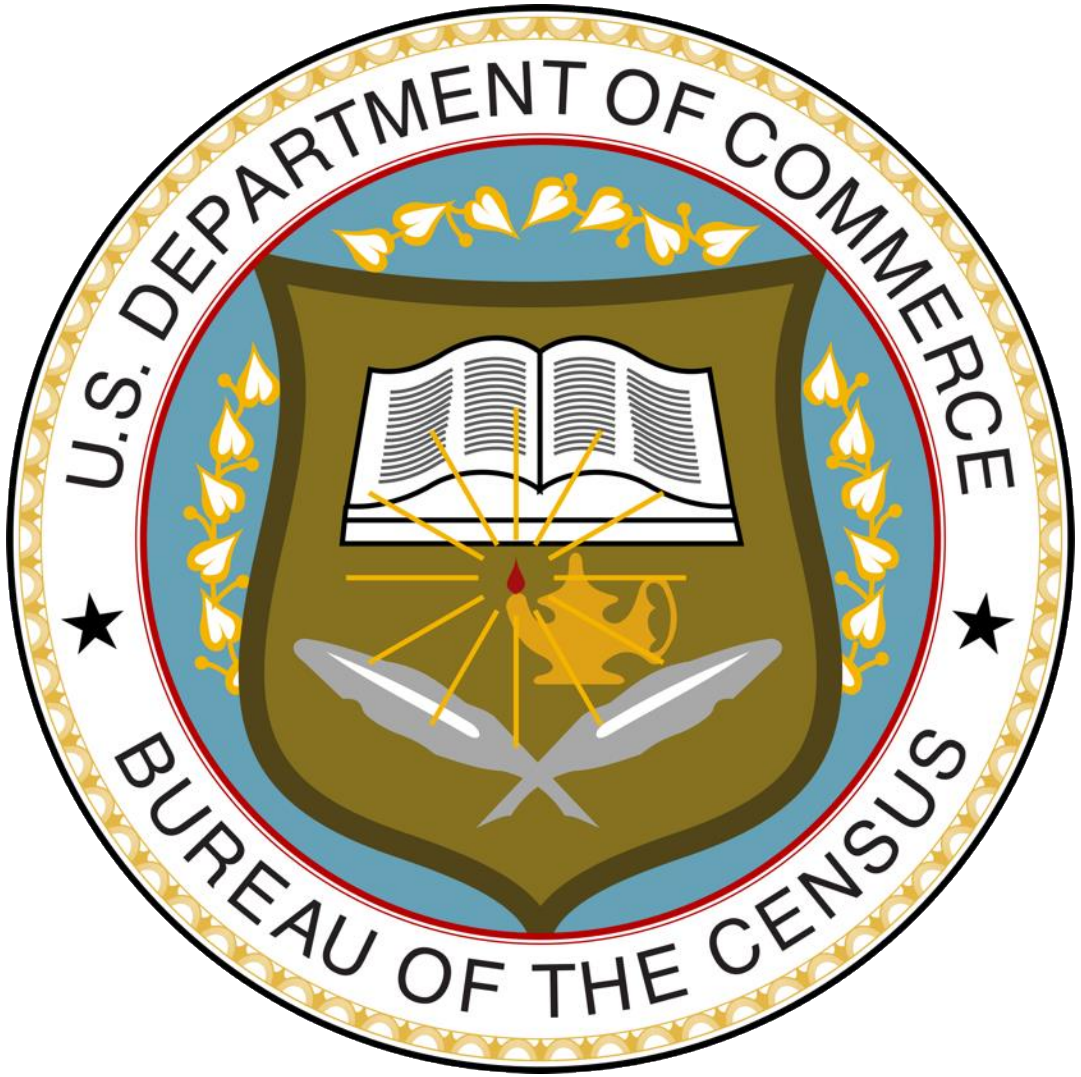
Open source research data repository software

figshare

github
SOCIAL CODING

Metadata


SMPTE
MPEG 7
MPEG 21
EBU P/META
IEEE
BBCi
SMEF
PBCore
Public Broadcasting Metadata Dictionary Project
METADATA
MARC 21
WGBH
TV-Anytime
Dublin Core Metadata Initiative
IMS
SCORM
Sharable Content Object Reference Model
Dewey Decimal Classification

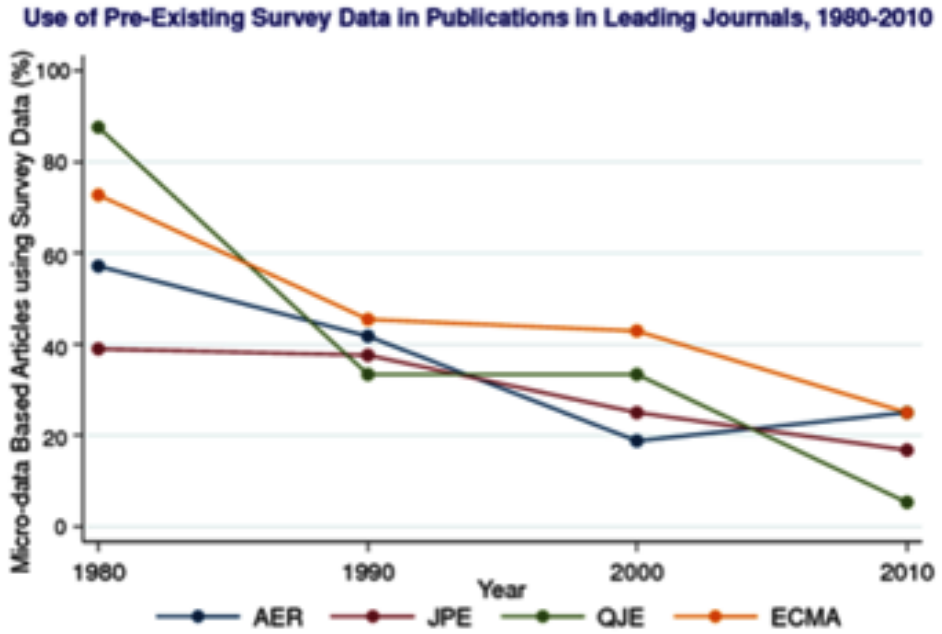


Research @ Census

Relevant characteristics of Census research

- Confidentiality and cloaking
- Tangled provenance
- Controlled environment
- Standardized computational tools (mostly)
- Standardized, mature metadata framework (mostly)

Increasing number of scholars pursuing research programs that mandate inherently identifiable data (e.g., geospatial relations, exact genome data, etc.)



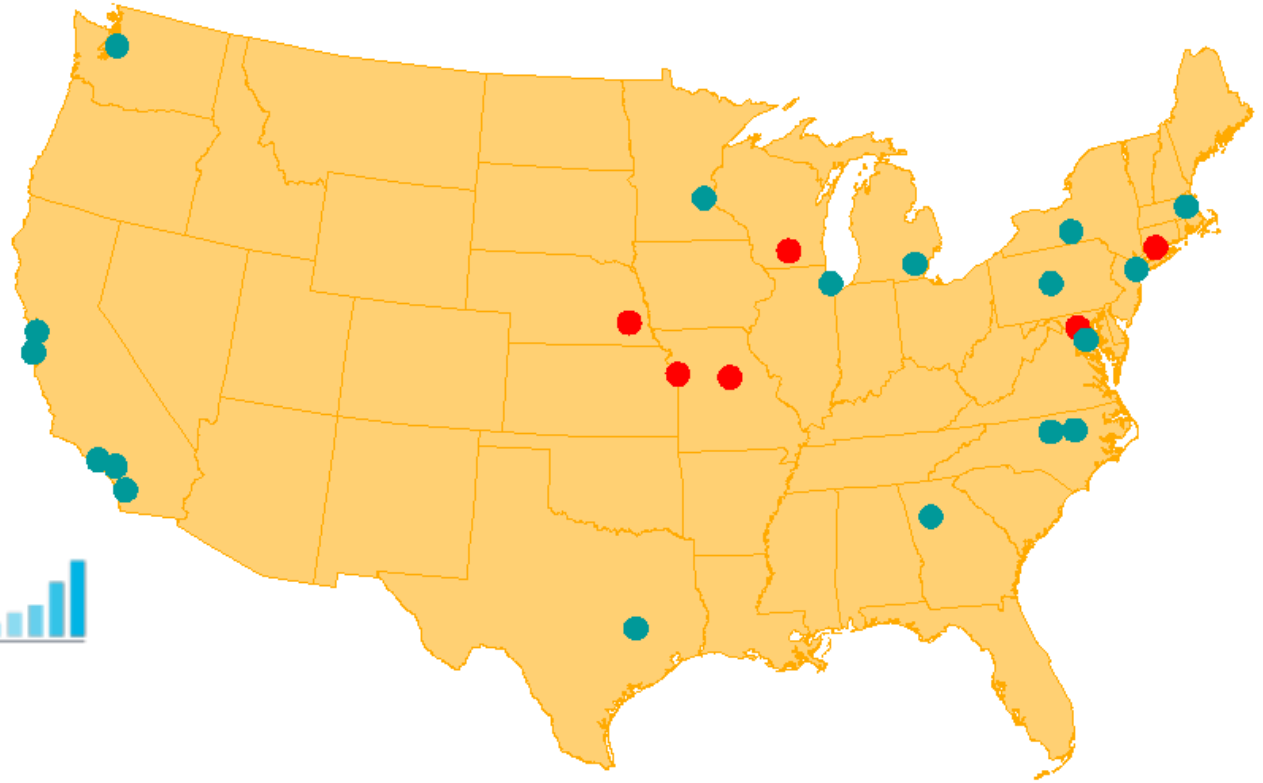
Chetty, R. (2012). The Transformative Potential of Administrative Data for Microeconometric Research. Retrieved from http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf

Problem is not limited to economics and social science

Many of the emerging “big data” applications come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results.

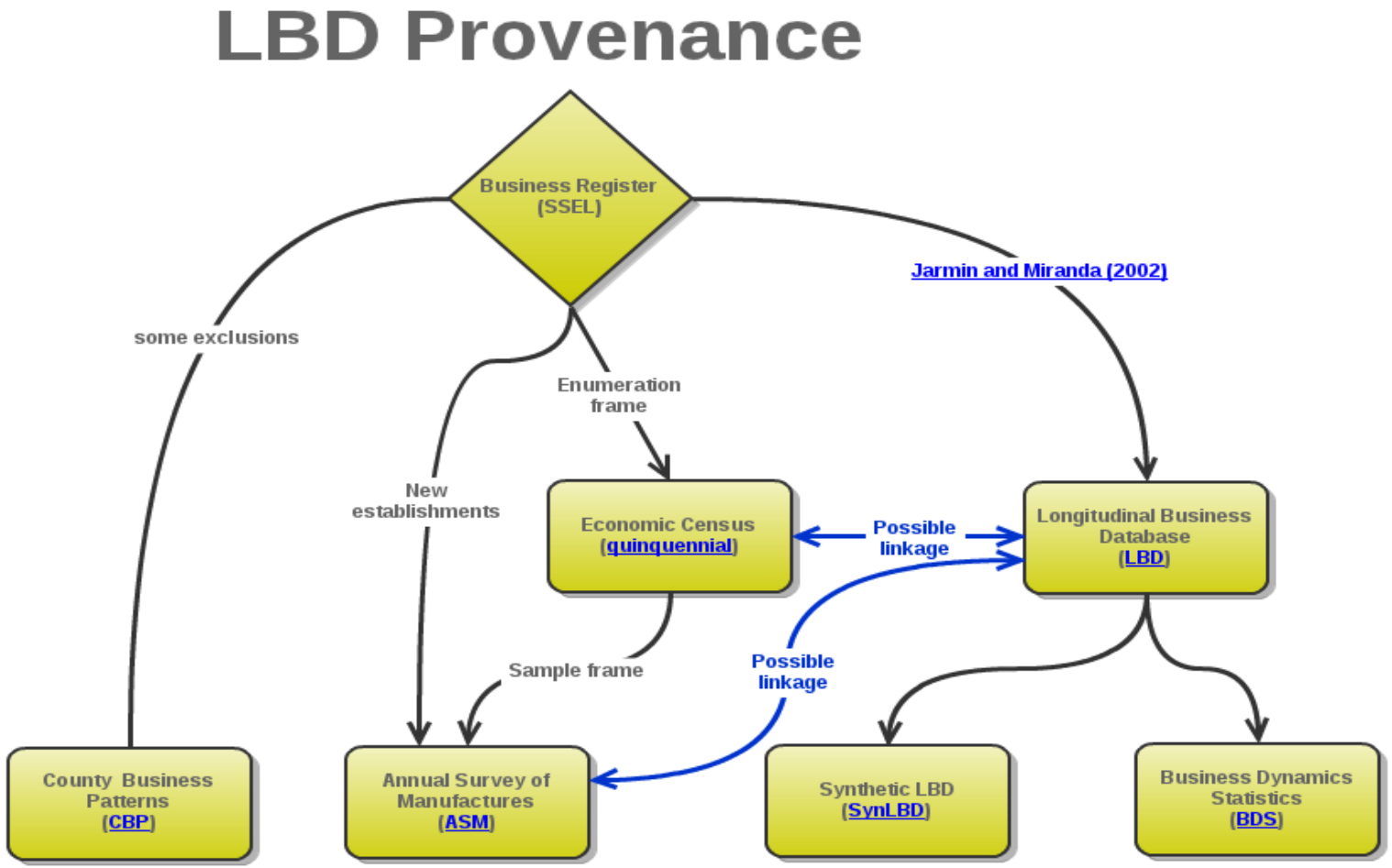
Huberman, Nature 482, 308 (16th February 2012)

Federal Statistical Research Data Centers



FSRDC

Restricted access data in the provenance chain complicates the curation and knowledge discovery process



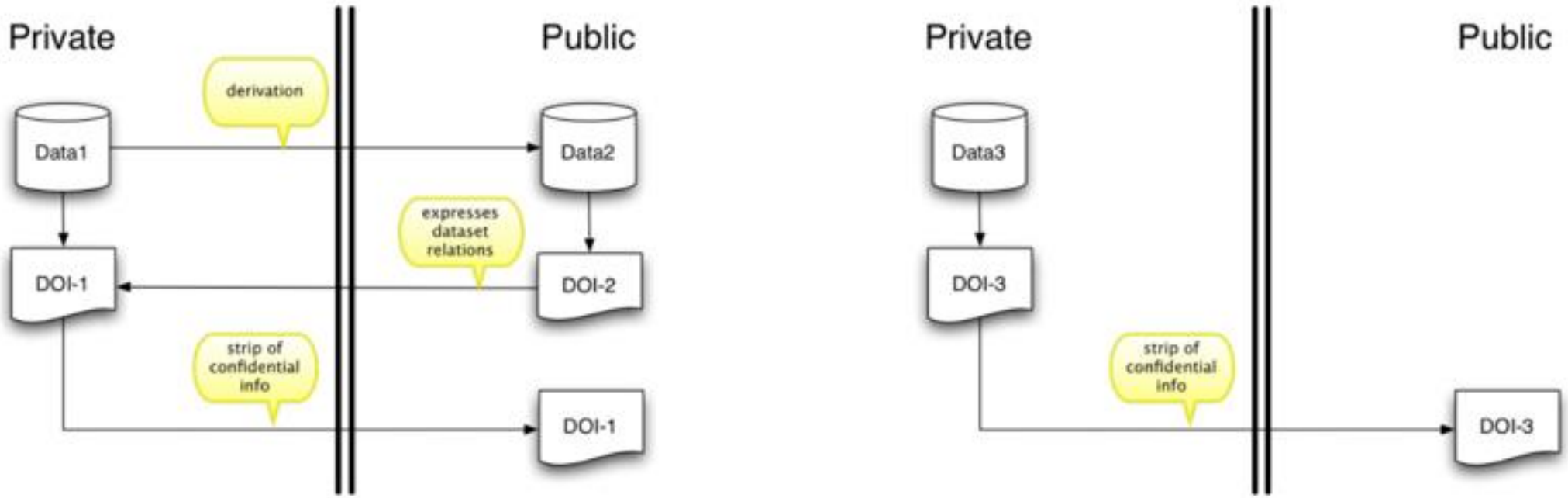


Figure 4. Two scenarios of confidential data and/or metadata. On the left, a data set exists in both a public and private (filtered and possibly enhanced) version, each with its own metadata, public and private, respectively; in addition, a filtered version of the private metadata is exposed publicly. On the right, only a single private data set exists with its own private metadata that is then filtered to the outside for the public use.

Summary of requirements

- curation of secure data sets
- Identification of all elements of research process
- selective hiding of data & metadata
- Rich provenance of data artifacts

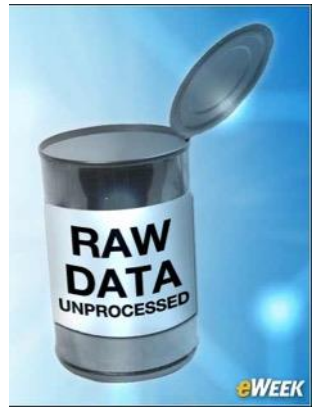
Need to leverage expertise of community to achieve these goals!!

Our Approach

- Rely on open standards, namely the Data Documentation Initiative (DDI) schema
- Leverage existing workflows
- Provide easy-to-use tools and interfaces to structured metadata
- Build infrastructure that enables data curators to leverage community-driven input to official documentation

Name everything

DATA CUISINE



University of California
CDL
California Digital Library

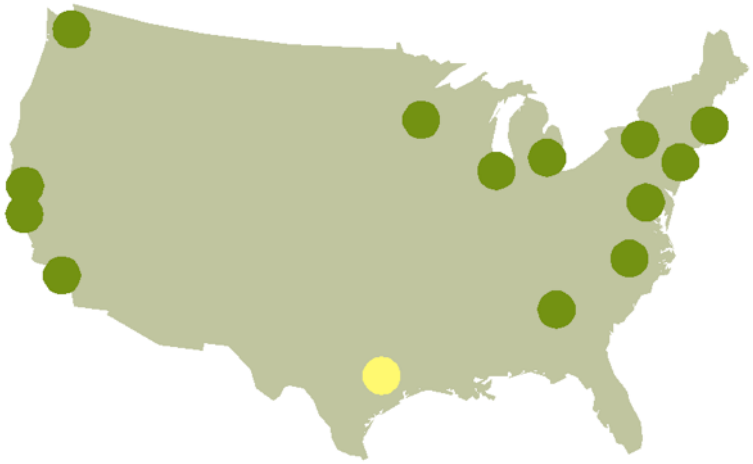
The ARK Identifier Scheme at Ten Years Old

doi

Handle.Net®

ORCID
Connecting Research and Researchers

Leverage existing workflows



Working Papers

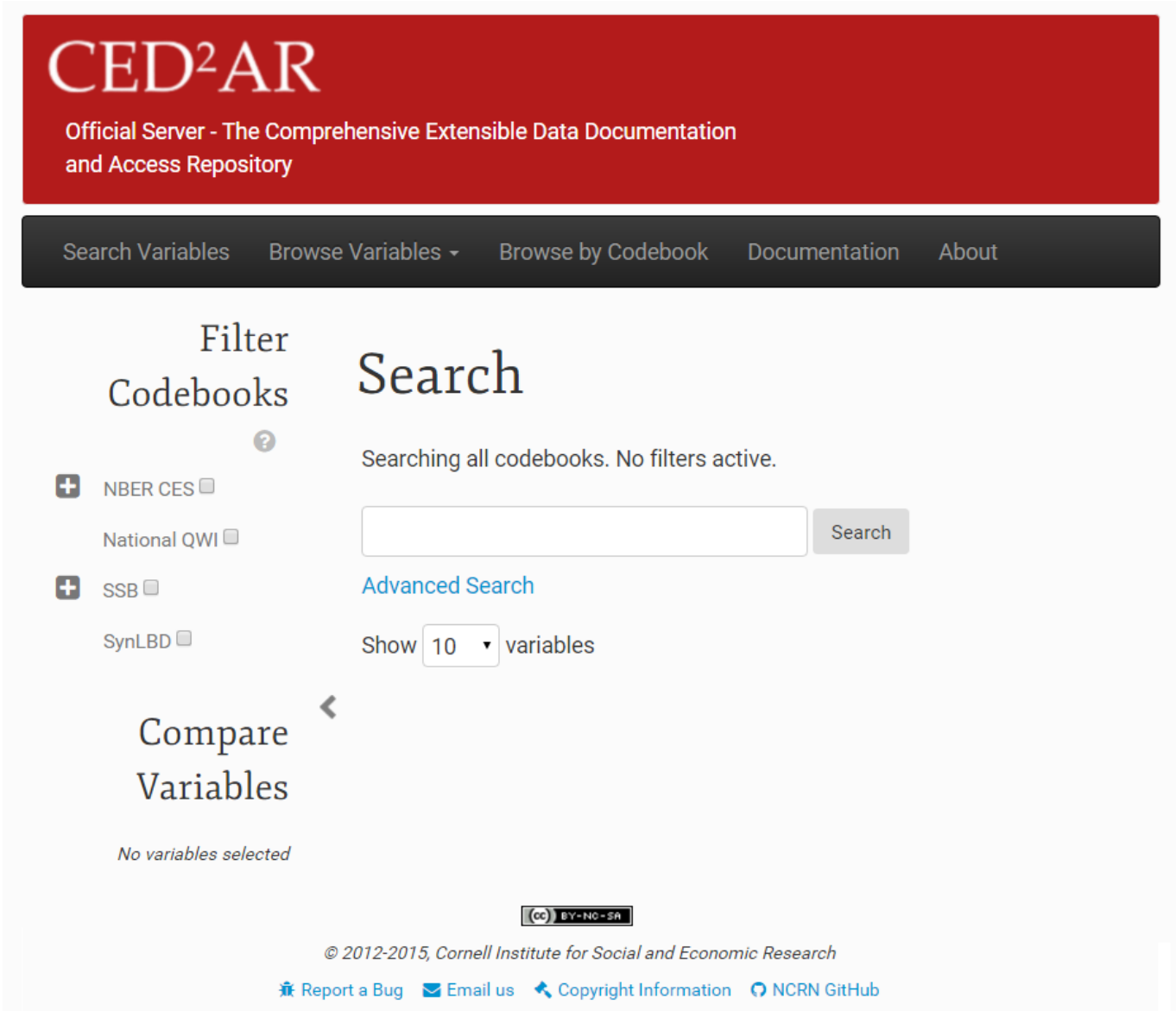
RePEc

CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

What is CED^{2}AR?

- Metadata curation software
- Designed for documenting discovering and providing access to existing datasets
- Support:
  - Confidentiality and cloaking
  - Provenance expression and visualization
  - Crowdsourcing of metadata
- Funded by NSF grant #1131848
- Online at www2.ncrn.cornell.edu/ced2ar-web


CED²AR
Official Server - The Comprehensive Extensible Data Documentation and Access Repository
Search Variables
Browse Variables
Browse by Codebook
Documentation
About
Filter Codebooks
NBER CES
National QWI
SSB
SynLBD
Compare Variables
No variables selected
Search
Searching all codebooks. No filters active.
Search
Advanced Search
Show 10 variables
© 2012-2015, Cornell Institute for Social and Economic Research
Report a Bug
Email us
Copyright Information
NCRN GitHub

Cloaking of data and metadata

Lagoze, C., Block, W., Williams, J., Abowd, J. M., & Vilhuber, L. (2013). Data Management of Confidential Data. Presented at the International Data Curation Conference, San Francisco, CA.

Expressing cloaking rules

```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obt    </confDec>
            <conditions>Confidential</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A3">
        <useStmt>
            <confDec>You're never gonna see this data.</confDec>
            <conditions>Need to know</conditions>
        </useStmt>
    </dataAccs>
</stdyDscr>
```

Variable level

```
<var ID="V1500" dcml="0" files="F3" intrvl="discrete" name="totfam_kids" access="A1">
    <location width="12"/>
    <labl>Total Number of Children in Family</labl>
    <valrng> [2 lines]
    <sumStat type="vald">1000</sumStat>
    <sumStat type="invd">0</sumStat>
    <varFormat schema="other" type="numeric"/>
</var>
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A2">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <valrng> [2 lines]
    <sumStat type="vald">240</sumStat>
    <sumStat type="invd">760</sumStat>
    <sumStat type="min">-278.739</sumStat>
    <sumStat type="max">39515.631</sumStat>
    <sumStat type="mean">1861.779</sumStat>
    <sumStat type="stdev">4015.033</sumStat>
    <varFormat schema="other" type="numeric"/>
</var>
```

Value level

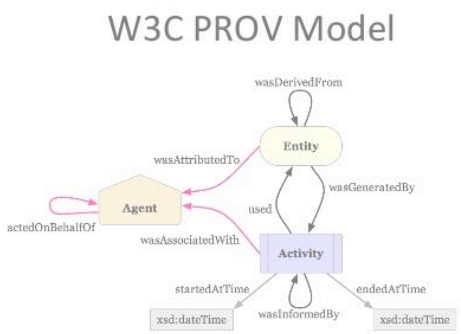
```
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A1">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <catgry>
        <catValu>0</catValu>
        <labl>5-25k</labl>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>25-75k</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>75-125k</labl>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>125-250k</labl>
    </catgry>
    <catgry access="A2">
        <catValu>4</catValu>
        <labl>250k+</labl>
    </catgry>
    <varFormat schema="other" type="numeric"/>
</var>
```

Provenance for social science data

Lagoze, C., Vilhuber, L., Williams, J., & Block, W. (2013). Encoding Provenance of Social Science Data: Integrating PROV with DDI. In *Proceedings of EDDI13 5th Annual European DDI User Conference*.

Combining DDI and W3C PROV models

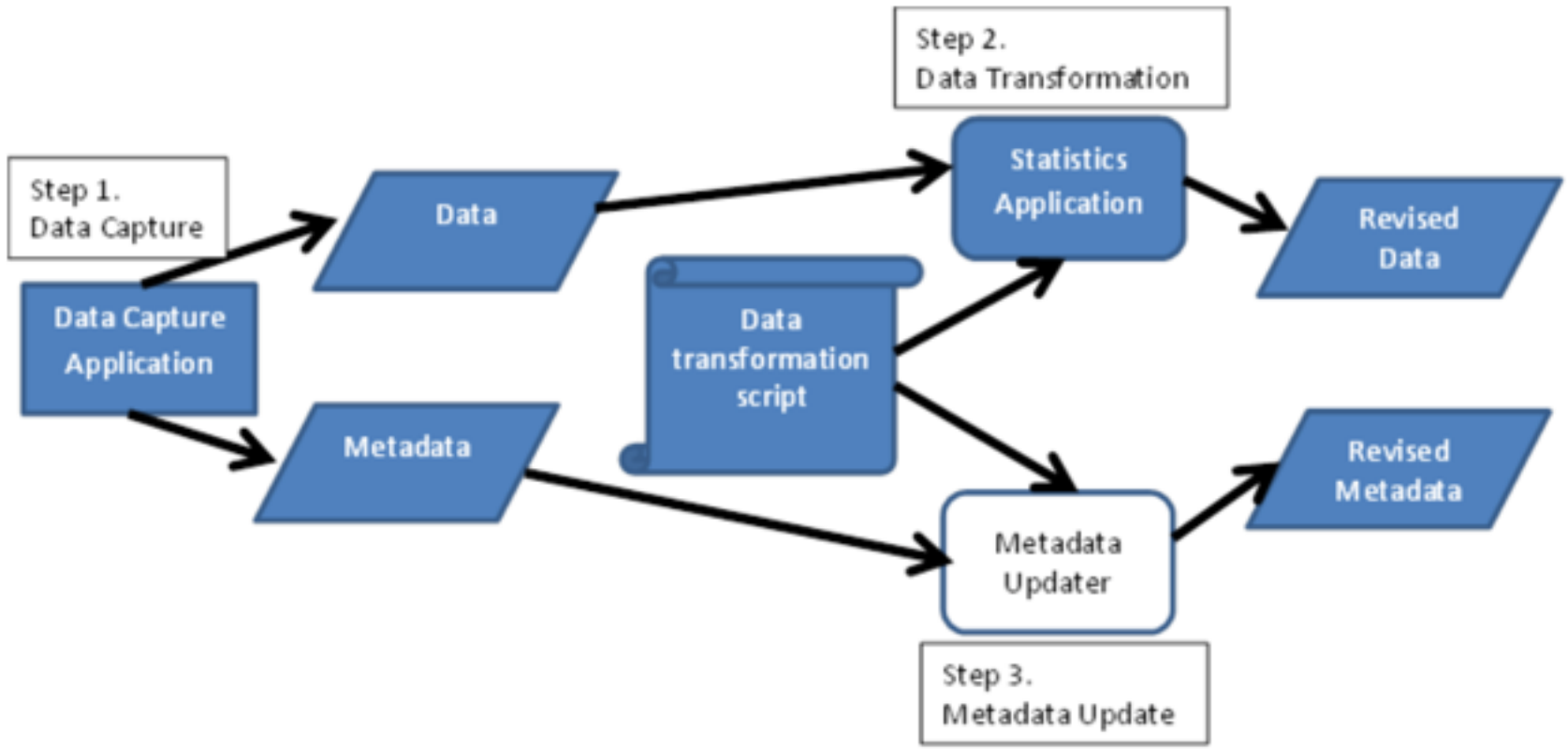
```
<Fragment xmlns="ddi:instance:3_2">
  <StudyUnit isUniversallyUnique="true" versionDate="2015-06-09T19:05:42.9389487Z" xmlns="ddi:
    <URN xmlns="ddi:reusable:3_2">urn:ddi:example.org:194d128c-93b2-4378-b626-2105edbdf561:4</
    <Agency xmlns="ddi:reusable:3_2">example.org</Agency>
    <ID xmlns="ddi:reusable:3_2">194d128c-93b2-4378-b626-2105edbdf561</ID>
    <Version xmlns="ddi:reusable:3_2">4</Version>
    <Citation xmlns="ddi:reusable:3_2">
      <Title>
        <String xml:lang="en-US">United States Census 2010</String>
      </Title>
      <description xml:lang="en-US" xmlns="http://purl.org/dc/elements/1.1/">United States Cen
    </Citation>
    <Abstract xmlns="ddi:reusable:3_2">
      <Content xml:lang="en-US">
        <div class="multi-audience" xmlns="http://www.w3.org/1999/xhtml">
          <div class="audience-default">The U.S. Census counts every resident in the United St
          <div class="audience-External Audience">A different abstract can be described for a
        </div>
      </Content>
    </Abstract>
    <UniverseReference xmlns="ddi:reusable:3_2">
      <Agency>example.org</Agency>
      <ID>b2200741-9232-4d97-91ed-3ff1fb11f818</ID>
      <Version>1</Version>
```



W3C PROV
http://www.w3.org/TR/prov-overview/
http://www.w3.org/TR/2013/REC-prov-o-20130430/

Curating and enhancing metadata – leveraging professionals and the crowd

Tools to capture data transformations from general purpose statistical analysis packages


Step 2.
Data Transformation
Step 1.
Data Capture
Data
Statistics
Application
Revised
Data
Data Capture
Application
Data
transformation
script
Metadata
Revised
Metadata
Metadata
Updater
Step 3.
Metadata Update

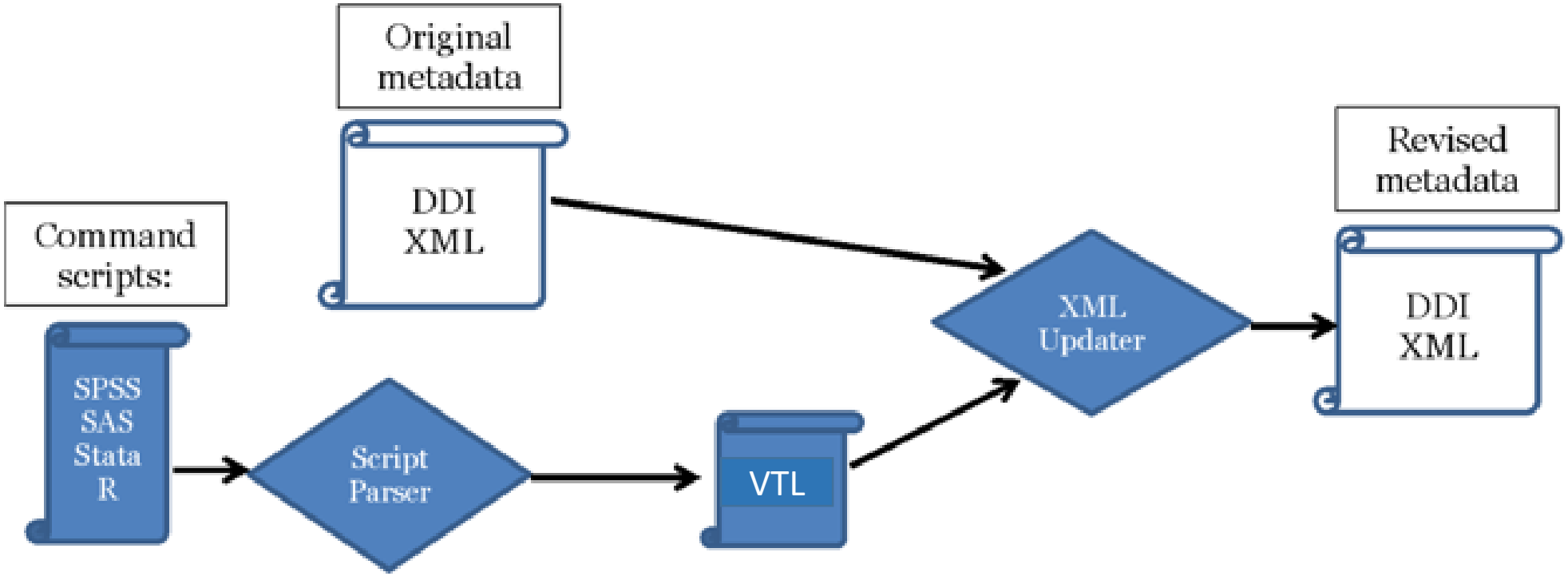


Figure 2. Metadata Capture

Conclusion

- Reproducibility is a general problem
- Reproducibility is a socio technical problem
- Solutions need to leverage community and general tools and practices