Research Data Preservation: Beyond Data Sharing and Open Science

Brett Currier, Bommae Kim, Cynthia Edwards, Courtney Butler

DLF Forum

October 23, 2017

Brett D. Currier

Digital Scholarship Librarian & Library Manager

Center for the Advancement of Data Research in
Economics

Federal Reserve Bank of Kansas City

The views expressed herein are solely those of
the author and do not necessarily reflect the
views of the Federal Reserve Bank of Kansas
City or the Federal Reserve System.

Defining Data

- Definition: Structured information used in quantitative analysis
 - Four Points of Preservation
 - Data Set
 - Extracted Raw Data
 - Intermediate Data
 - Final Model Data
-

Where does data come from?

Primary Data

- Collected data
 - Health Science Studies
- Created data
 - Systematic Reviews

Secondary Data

- Publicly available data
 - Census
- FOIA data
- Purchased data
- Restricted use data
 - Research Data Centers

Restrictions

On Raw Data

- Limits on internal access
- Limits on sharing
- Limits on preservation

On aggregated or computed data

(Intermediate or Final Data Sets)

- Limits on internal access
 - Limits on sharing
 - Limits on preservation
-

Where do restrictions come from?

Primary Data

- IRB Study
- Non-disclosure agreement to participants
- End User License Agreement
- Site License
- State Law
- Federal Law

Secondary Data

- End User License Agreement
- Site License
- State Law
- Federal Law
- Formal Contracts

Preservation

- 10 year half life
 - 4 different updates to Microsoft Office in 10 years
 - Restrictions are barriers to preservation
-

Determine the Reason for the Underlying Restriction

- Confidential Information
 - IRB Information
 - Research Data Centers
- Data Integrity
- Industry Leader
 - ICPSR: Standard Agreement
- Proprietary Data
- Market harm through competition

Questions

- Does the underlying restriction
 - Bar public dissemination?
 - Bar private preservation?
- Does it matter?
- Is the vendor or participant willing to given consent?
- To what extent is the library or institution willing to mediate or facilitate those conversations?
- Can the data be aggregated or computed enough from the underlying raw data to satisfy concerns?

Questions?