Developing software to prepare social science research data and code for sharing and preservation

Limor Peer

Institution for Social and Policy Studies, Yale University
Data sharing
Using other people’s data...

“The most commonly reported problems associated with replication attempts were the lack of data and code, followed by insufficient documentation.”

Usable data:
Intelligently open
Independently understandable

Introduction

Unusable data = lost data

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

IF:
Shared (and/or preserved) data may not be usable

THEN:
Make data usable = data curation
Project to develop curation software

• Background
• Requirements
• The software
• The architecture at Yale

Data curation
Active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.

-- The University of Illinois' Graduate School of Library and Information Science
ISPS Data Archive

Research

The ISPS KnowledgeBase is the gateway to all ISPS data projects, and publications. It is an integrated database which provides a one-stop-shop for ISPS-related research products.

Search the KnowledgeBase or browse recent additions.

Yale ISPS KnowledgeBase

- Terms of use
- About the ISPS data archive
- AUTHOR
- AREA OF STUDY
- DISCIPLINE
- YEAR
- LOCATION
- KEYWORDS
- RESEARCH DESIGN

SEARCH ISPS

Research Funding

ISPS invites proposals for important and well-crafted field experiments in the social sciences and related policy issues. Field experiments are fully-randomized research designs in which observations found in a naturalistic setting – voters, patients, welfare recipients, community organizations, government entities, and the like – are assigned to treatment and control conditions (see more here).

- Apply for a field experiment grant
- Additional funding opportunities

ISPS Working Paper Series

ISPS advances interdisciplinary research in the social sciences that aims to shape public policy and inform democratic deliberation. The ISPS network includes scholars and students from many departments in the Faculty of Arts and Sciences and from Yale’s graduate and professional schools as well as select experts from other institutions. The ISPS Working Paper Series provides a platform for ISPS affiliates to make
ISPS Data Archive

Data

Terms of use

About the ISPS data archive

AUTHOR
- Any -

AREA OF STUDY
- Any -

DISCIPLINE
- Any -

YEAR
- Any -

LOCATION
- Any -

KEYWORDS
- Any -

RESEARCH DESIGN
- Any -

Search

ABOUT OUR DATA

On this page you will find the ISPS Data Archive. The majority of digital content in the ISPS Data Archive currently consists of social science research data from experiments, program files with the code for analyzing these data, requisite documentation to use and understand the data, and associated files. Access to the ISPS Data Archive is provided at no cost and is granted for scholarship and research purposes only. When possible, Data is linked to Projects and Publications, via the ISPS KnowledgeBase.

More about the ISPS Data Archive.

ISPS DATA ARCHIVE: TERMS OF USE

By using, contributing, and/or downloading files associated with scholarly studies available on the ISPS Data Archive, you agree to these terms and conditions. Please read the ISPS Data Archive Terms of Use.
Data Quality Review

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

Two Research Organizations

The two organizations have in common...

• Similar content: Data from randomized controlled trials in the social sciences
• Similar approach to data sharing and preservation: Focus on replication, review data and code pre-publication

Cross-fertilization...

• Build on ISPS Data Archive curation standards and practices
• Maintain key aspects of ISPS UI such as linked publications, data, and code
• Build on IPA ability to prepare data earlier in the lifecycle (e.g., pre analysis)
• Allow IPA network to access software from distributed research sites
ISPS and IPA Requirements

• Curation workflow management (dashboard)
• Track changes to files (provenance)
• Integrate metadata production with data and code review and cleaning
• Preservation metadata and formats
• Secure storage and access
• Smooth transition to public dissemination of content
• Preference for open source solutions
Curator software: Making data usable

A software platform that leverages the DDI Lifecycle and structures the curation workflow, including checking data for confidentiality and completeness, creating preservation formats, and reviewing and verifying code.

Experts in social science metadata involved in DDI development.
Features

• Web-based
• Built on DDI 3.2
• Open Source
• Builds on Existing Tools
A Field Experiment on Legislators' Home Styles: Service versus Policy
# A Field Experiment on Legislators' Home Styles: Service versus Policy

<table>
<thead>
<tr>
<th>File</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butler_et_al_JP_2012_metadata.pdf</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Hyde_Marthov_IO_2014_observers.csv</td>
<td>Data</td>
<td>Accepted</td>
</tr>
<tr>
<td>Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.dta</td>
<td>Data</td>
<td>Accepted</td>
</tr>
<tr>
<td>Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.xml</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Butler_et_al_JP_2012_JOP_Service_v_Policy_Limited_Do_File.do</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.csv</td>
<td>Data</td>
<td>Accepted</td>
</tr>
<tr>
<td>Butler_et_al_JP_2012_JOP_Service_v_Policy_Limited_R_File.R</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>File</td>
<td>Type</td>
<td>Status</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>Gerber.et_al_JP_2011_README.pdf</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Gerber_et_al_JP_2011_02_subC3_performmturk_analysis.do</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Gerber.et_al_JP_2011_nturk_replication.dta</td>
<td>Data</td>
<td>Pending</td>
</tr>
</tbody>
</table>

- **Add or Update Files**
- **Download All Files**

**Check Missing Labels**
- Review Observation Count
- Compare Questionnaire, Codebook, and Data in Data File
- Check for Personally-Identifiable Information (PI) in Data File
- Identify Potential Errors in Data File
<table>
<thead>
<tr>
<th>File</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gerber_et_al_JP_2011_README.pdf</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Gerber_et_al_JP_2011_C2_sub03_performmturk_analysis.db</td>
<td>Other</td>
<td>Accepted</td>
</tr>
<tr>
<td>Gerber_et_al_JP_2011_nturk_replication.dta</td>
<td>Data</td>
<td>Pending</td>
</tr>
<tr>
<td>Title</td>
<td>Collection</td>
<td>Processing</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Field Experiment on Legislators' Home Styles: Service versus Policy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>An Experiment Testing the Relative Effectiveness of Encouraging Voter Participation by Inducing Feelings of Pride or Shame</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessing the Stability of Psychological and Political Survey Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Technical components & support at Yale

## ITS

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hardware</strong></td>
<td>Windows Server (VM), 32GB RAM minimum (8 Cores), 100GB local disk for OS, applications and swap files</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Colectica suite of tools, statistical software, integrated APIs</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>RSS start at 500GB, read/write/no-execute access to one or more directories</td>
</tr>
<tr>
<td><strong>Application hosting</strong></td>
<td>WCF application and ASP.NET MVC web application on IIS, plus a SQL Server database (10GB), a Windows Service</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>Federated identification</td>
</tr>
</tbody>
</table>

## Library

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Long-term preservation</strong></td>
<td>Fedora Commons / Hydra</td>
</tr>
<tr>
<td><strong>Discovery</strong></td>
<td>Blacklight</td>
</tr>
<tr>
<td><strong>Persistent links</strong></td>
<td>Handle service (ODAI)</td>
</tr>
</tbody>
</table>
(Target) Timeline

Project Kickoff – February 2014
Development Plan – March to April
Design + Base Platform and Basic Workflow development – May to October
Full Workflow Development – November to December
Ongoing development and maintenance – January 2015+
Thank you!

limor.peer@yale.edu
@l_peer

In collaboration with:

**Innovations for Poverty Action**: Niall Keleher, Stephanie Wykstra

**Digital Lifecycle Research and consulting**: Ann Green

**Colectica** software company: Jeremy Iverson, Dan Smith

**Yale ITS Academic IT / Research Services**: Kiran Keshav, Themba Flowers, Paul Gluhosky

**Yale Library IT**: Michael Dula, Mike Friscia, Eric James

**Yale Library CSSSI**: Michelle Hudson, Jill Parchuck

and Yale ODAI and Office of General Counsel
2014 YALE TECHNOLOGY SUMMIT

Friday, October 31, 2014
Evans Hall, Yale School of Management

#YaleTechSummit2014

Presented by Yale Information Technology Services