AN OPEN SOURCE, DDI-BASED DATA CURATION SYSTEM FOR SOCIAL SCIENCE DATA
Two Partners, a Consultant, and a Software Developer

Digital Lifecycle Research & Consulting
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.

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

Featured Books By ISPS Faculty
Two Research Organizations

Institution for Social and Policy Studies (Yale)
- Data preparation at end of research project
- Replication
- Field Experiments
- Linked publications, data, and code

Innovations for Poverty Action
- Data preparation before analysis and at end of research project
- Project hosting from distributed research sites
- Lifecycle data management
Why make something new?

The Need: Curation Management
Data Quality Review

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
ISPS Data Pipeline

Processing Space:
1. Identify files
2. Obtain files
3. Review files
4. Archive original files
5. Process files
6. Review files

Archive Space:
7. Archive processed files

Web:
8. Publish files
9. Announce
Processing Steps

1. Assign staff to study and files
2. Move original files to Archive space
3. Make copies of processed files and move to collaborative space
4. Identify related publications and projects
5. Rename all copied files for public dissemination according to ISPS Data Archive naming conventions
6. Check and complete variable-level metadata for each data file
7. Compare variable information, check for additional variables and recoded variables, check variable/value labels
8. Check all files for confidential and other sensitive information
9. Run the statistical code and check against published results
10. Re-write statistical code in R and check replication
11. Communicate with PI as needed
12. Create new DDI-XML file with variable-level information
13. Create additional files by converting to readable formats (e.g., ASCII, PDF)
14. Update study- and file-level metadata record
15. Update tracking documents: process record / general study database / status document
Solution: Build a Curation Management System to Work with Existing Tools
Neat Features

- Built on DDI 3.2
- Web-based
- Open Source
User Roles

- Depositor
- Curator
- Administrator
- Machines
- Researchers
Goal: Ensure no missing variable labels

Current Approach
- Use Stata to open .dta file
- Manually scan for missing labels
- Use Stata to edit and save new copy of .dta file
- Use Excel to make changes to metadata and “process record”
Goal: Ensure no missing variable labels

New Approach

- Curator opens Web application
- Curator sees a list of variables with missing labels
- Curator adds labels as appropriate
- The system logs this information and generates a new .dta file
Data Migration

Curation System

Yale ISPS

IPA

Other Data Sources
## Timeline

<table>
<thead>
<tr>
<th>Now</th>
<th>January 2015</th>
<th>2015+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development</td>
<td>Packaging</td>
<td>Ongoing Development</td>
</tr>
<tr>
<td>Testing</td>
<td></td>
<td>Maintenance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wider Adoption</td>
</tr>
</tbody>
</table>
Demonstration
Thank you

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Organization</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann Green</td>
<td>Independent Consultant</td>
<td><a href="mailto:green.ann@gmail.com">green.ann@gmail.com</a></td>
</tr>
<tr>
<td>Jeremy Iverson</td>
<td>Colectica</td>
<td><a href="mailto:jeremy@colectica.com">jeremy@colectica.com</a></td>
</tr>
<tr>
<td>Niall Keleher</td>
<td>Innovations for Poverty Action</td>
<td><a href="mailto:nkeleher@poverty-action.org">nkeleher@poverty-action.org</a></td>
</tr>
<tr>
<td>Limor Peer</td>
<td>Yale University</td>
<td><a href="mailto:limor.peer@yale.edu">limor.peer@yale.edu</a></td>
</tr>
<tr>
<td>Dan Smith</td>
<td>Colectica</td>
<td><a href="mailto:dan@colectica.com">dan@colectica.com</a></td>
</tr>
<tr>
<td>Stephanie Wykstra</td>
<td>Innovations for Poverty Action</td>
<td><a href="mailto:swykstra@poverty-action.org">swykstra@poverty-action.org</a></td>
</tr>
</tbody>
</table>