YARD Work: Data Archiving and Publishing at Yale

Limor Peer, Joshua Dull, Themba Flowers

Yale Digital Conference
June 15, 2018
Yale Institution for Social and Policy Studies
isps.yale.edu
ISPS Data Archive
Curation Tool: Yale Application for Research Data (YARD)
YARD-Drupal integration
ISPS was founded in 1968 as an interdisciplinary center to support social science and public policy research at Yale University.
Study Shows Op-Eds Change Minds

Alex Coppock’s new research in the Quarterly Journal of Political Science

READ MORE ➤
An open access digital collection of social science experimental data, metadata, code, and associated files produced by ISPS researchers, for the purpose of replication of research findings, further analysis, and teaching.


http://dx.doi.org/10.2218/ijdc.v7i1.222
ISPS Data Archive

Since 2011
Open access
Creative Commons license
Website integration

Specialized community
Experimental design
Quantitative data

90 studies
1,400 files
15 GB

https://isps.yale.edu/research/data
The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author.

Common Replication Problems

- Insufficient documentation
- Missing variables
- Deviations in number of observations
- Unavailable software extensions
- Omitted code
- Incompatible datasets
Curating for Reproducibility

Data Curation: Maintaining, preserving and adding value to digital research data throughout its lifecycle... [which] reduces threats to their long-term research value and mitigates the risk of digital obsolescence.

Digital Curation Center http://www.dcc.ac.uk/digital-curation/what-digital-curation

Reproducibility: Calculation of quantitative scientific results by independent scientists using the original datasets and methods.

Curating for Reproducibility

Data Quality Review

- Assign persistent identifier
- Create study citation and study-level metadata record
- Record file size details
- Check for presence of all files
- Verify content of files matches expected format
- Create non-proprietary versions of files
- Implement migration strategy for file formats
Curating for Reproducibility

Data Quality Review

✓ Confirm presence of comprehensive descriptive information necessary for informed reuse
  • Data definitions
  • Variable construction
  • Methodology
  • Sampling information
  • Original data source citation
  • Analysis software version

✓ Link to related research products
Curating for Reproducibility

Data Quality Review

✓ Check for undocumented variable and value information
✓ Examine data for inconsistencies and errors
  • Discrepancies in number of observations
  • Out-of-range or wild codes
  • Undefined null values
✓ Review data for confidentiality issues
Curating for Reproducibility

Data Quality Review

- Convert absolute file paths to relative file paths
- Check code for presence of non-executable comments that document analysis processes
- Identify packages required to execute code
- Execute code to ensure code is error-free
- Compare code output to findings presented in article
Typical ISPS 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

... and now share the data!
Curating for Reproducibility at ISPS

YARD: Yale Application for Research Data

A new workflow tool that allows Depositors, Curators, and Administrators to submit, review, process, and publish data within one system.

The software structures the curation and review workflow and all actions are recorded in the system.

The tool integrates and captures DDI metadata production with data and code review and cleaning.

Processed data packages are directed to pre-specified destinations.
Curation Tool: YARD

Yale Application for Research Data

Log in

Log in to the ISPS Data Curation Tool with your username and password.

Don't have an ISPS Data Curation Tool account?
Create an account.

Email

Password

Remember me

Log in

Forgot your password?

Documentation: https://docs.colectica.com/curation/
Curation Tool: YARD

Three roles:
• Depositor: Any User
• Curator: Trained Archive Staff
• Administrator/Approver: Archive Manager
Curation Tool: YARD

Data Deposit

Create New Catalog Record

General

Title *
Authors *
Add or remove authors

Description *
Number
Keywords *
Separate multiple outcome measures with a comma.

Citation

Funding *

Access

Access Statement *
Not Selected
Curation Tool: YARD

Data Deposit

Research Study 2

Methods

- Research Design: Natural experiment
- Mode of Data Collection: Interview: Face to Face
- Field Dates: 2017-05-19
- Study Time Period: 2013-05-08
- Location: Africa
- Unit of Observation: Individual
- Sample Size: 5000
- Inclusion/Exclusion Criteria: None
- Randomization Procedure: None
Curation Tool: YARD

Data Deposit
Curation Tool: YARD

Data Deposit

Submit for Curation

Your record is ready for curation. Click the button below to submit it.

Deposit Agreement

ISPS Data Archive Data Deposit Agreement

Updated October 5, 2015

Under General Counsel review.

Terms of Data Deposit Agreement

Your decision to deposit data in the ISPS Data Archive requires that you accept the terms of this Data Deposit Agreement. This agreement grants permission to Yale University, on behalf of ISPS, to (a) archive files associated with scholarly studies, and (b) to distribute these files for the purpose of replication (hereafter, Replication Files).

Specifically, by using the ISPS Data Archive, you accept the terms of this agreement, and you (or your proxy, on your behalf) agree to the following (please check each item if you agree):

(1) You accept and agree to abide by ISPS Terms of Use <link to http://isps.yale.edu/research/data/terms-of-use>.
(2) You agree that you will comply with all applicable laws and licenses.
(3) If you are not the data owner, you represent that you have obtained the unrestricted permission of the owner to grant Yale University, on behalf of ISPS, the rights required by this license, that you will inform the data owner that ISPS administrators have access to all content, for purposes of maintaining the system and assisting users, in accordance with the policies specified here (ISPS does not make data available by other means such as FTP removable media, etc.), and that the owner of the material will be clearly identified and acknowledged within the text or content of the submission.

Important: The ISPS Data Archive expects all data contributors to comply with applicable copyright laws. Use of third-party copyrighted material without permission or that does not fall under one of the statutory exceptions, including but not limited to fair use, is illegal, unethical, and exposes Yale University to significant financial liability.
Curation Tool: YARD

Data Deposit

1. Create an account
2. Create new record
3. Fill in ‘Study Level’ metadata
4. Upload files...
5. Submit for curation

Storage@Yale

Virus Check

Collected Repository (DDI)

data/ingest/UUID
data/processing/UUID

Bag created
data/archive/UUID-
ingest.zip
Curation Tool: YARD

Data Curation

[Image of the YARD tool interface with tasks and statuses]
Curation Tool: YARD

Data Curation

POQ_PublicReplicationDatafile.dta

- Variables
- Notes
- Review

weight
case weight

yal130
voting private matter

Frequencies

<table>
<thead>
<tr>
<th>Value</th>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>democracy works best when people treat their vote choices as personal matters</td>
<td>556</td>
</tr>
</tbody>
</table>

Depositor
Curator
Approver
Curation Tool: YARD

Data Curation

Request Publication

When curation is complete, you can request publication.

Please complete the following fields before requesting publication.

- Incomplete task for 02_PerformAnalysis.dta: Confirm Code Executes
- Incomplete task for testdattest.docx: Create Preservation Format
- Incomplete task for POQ_PublicReplicationDatafile.dta: Check Missing Labels
- Incomplete task for POQ_PublicReplicationDatafile.dta: Identity Potential Errors in Data File
- Incomplete task for POQ_PublicReplicationDatafile.dta: Check for Personally-Identifiable Information (PII) in Data File
- Incomplete task for 02_PerformAnalysis.dta: Confirm Code Replicates Reported Results
- Incomplete task for POQ_PublicReplicationDatafile.dta: Compare Questionnaire, Codebook, and Data in Data File
- 02_PerformAnalysis.dta: Software
- testdattest.docx: Software
- POQ_PublicReplicationDatafile.dta: Data Type
- testdattest.pdf: Software
- README.txt: Software
Curation Tool: YARD

Data Publication

Research Study 2

Publication Requested | Curator | Approver

General | Methods | Files | Review | Publish and Archive

Publish and Archive

- Approve Publication
- Reject Publication

Depositor
Curator
Approver
Curation Tool: YARD

Data Curation & Publication

1. Assigned to record
2. Complete curation tasks
3. Apply metadata updates
4. Request Publication
5. Review and approve publication

Preservation Formats
Handles & Checksum
DDI metadata file created
Bag created
Storage@Yale
Drupal Integration

XML Feed
Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

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.

Author(s): Alexander Coppock

Suggested citation:
http://hdl.handle.net/10079/zw3r2f5, ISPS Data Archive.

ISPS ID: D149

Related publications:
Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

Keyword(s): List experiment
Election Polling
Research design: Survey experiment
Data type: Survey/Interview (e.g., ANES)
Drupal Integration

File Level

<table>
<thead>
<tr>
<th>DATA FILE NUMBER</th>
<th>DESCRIPTION</th>
<th>FILE FORMAT</th>
<th>SIZE</th>
<th>FILE URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>D149F01</td>
<td>ReadMe file</td>
<td>.txt</td>
<td>519</td>
<td>Download file</td>
</tr>
<tr>
<td>D149F02</td>
<td>Data file</td>
<td>.csv</td>
<td>481797</td>
<td>Download file</td>
</tr>
<tr>
<td>D149F03</td>
<td>Program file</td>
<td>.r</td>
<td>11384</td>
<td>Download file</td>
</tr>
<tr>
<td>D149F04</td>
<td>Metadata (DDI 3.2)</td>
<td>.xml</td>
<td>26771</td>
<td>Download file</td>
</tr>
</tbody>
</table>
## Drupal Integration

### Required Modules

<table>
<thead>
<tr>
<th>Module</th>
<th>Version</th>
<th>Description</th>
<th>Dependencies</th>
<th>Required by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field Collection Feeds</td>
<td>7.x-1.0- alpha4</td>
<td>Provides feeds integration for field collection.</td>
<td>Requires: Field collection (enabled), Entity API (enabled), Feeds (enabled), Chaos tools (enabled), Job Scheduler (enabled)</td>
<td>Drupal (Feeds is currently using this module for one or more importers)</td>
</tr>
<tr>
<td>Field Validation</td>
<td>7.x-2.6</td>
<td>Add validation rules to fields.</td>
<td>Requires: Field (enabled), Field SQL storage (enabled), Token (enabled), Chaos tools (enabled)</td>
<td>Field validation extras (disabled), Field Validation UI (enabled)</td>
</tr>
<tr>
<td>Field validation extras</td>
<td>7.x-2.6</td>
<td>Extra validators for Field validation 7.x-2.x.</td>
<td>Requires: Field Validation (enabled), Field (enabled), Field SQL storage (enabled), Token (enabled), Chaos tools (enabled)</td>
<td></td>
</tr>
<tr>
<td>Field Validation UI</td>
<td>7.x-2.6</td>
<td>UI for Field Validation.</td>
<td>Requires: Field Validation (enabled), Field (enabled), Field SQL storage (enabled), Token (enabled), Chaos tools (enabled)</td>
<td></td>
</tr>
<tr>
<td>Pathauto</td>
<td>7.x-1.3</td>
<td>Provides a mechanism for modules to automatically generate aliases for the content they manage.</td>
<td>Requires: Path (enabled), Token (enabled)</td>
<td>Required by: Cultura (disabled), Easy Breadcrumb (disabled), ELTI General (disabled), Yale Common (enabled), projects (enabled)</td>
</tr>
</tbody>
</table>
Drupal Integration

Feeds Importers

Create one or more Feed importers for pulling content into Drupal. You can use these importers from the Import page or - if you attach them to a content type - simply by creating a node from that content type.

Add importer | Import importer

<table>
<thead>
<tr>
<th>NAME</th>
<th>DESCRIPTION</th>
<th>ATTACHED TO</th>
<th>STATUS</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>curation_data</td>
<td>[name]</td>
<td>Normal</td>
<td>Edit</td>
<td>Export</td>
</tr>
<tr>
<td>curation_data_files</td>
<td>[name]</td>
<td>Normal</td>
<td>Edit</td>
<td>Export</td>
</tr>
<tr>
<td>curation_keywords</td>
<td>[name]</td>
<td>Normal</td>
<td>Edit</td>
<td>Export</td>
</tr>
<tr>
<td>curation_location</td>
<td>[name]</td>
<td>Normal</td>
<td>Edit</td>
<td>Export</td>
</tr>
<tr>
<td>curation_research_design</td>
<td>[name]</td>
<td>Normal</td>
<td>Edit</td>
<td>Export</td>
</tr>
</tbody>
</table>
Drupal Integration

Feeds Importers

```
$feeds_importer = new stdClass();
$feeds_importer->disabled = FALSE; /* Edit this to true to make a default feeds_importer disabled initially */
$feeds_importer->api_version = 1;
$feeds_importer->id = 'curation_data';
$feeds_importer->config = array(
    'name' => 'curation_data',
    'description' => '',
    'fetcher' => array(
        'plugin_key' => 'FeedsHTTPFetcher',
        'config' => array()
    )
);
```
Drupal Integration

Mapping to Drupal

```php
array('processor' => array(
    'plugin_key' => 'FeedsNodeProcessor',
    'config' => array(
        'expire' => '-1',
        'author' => 0,
        'authorize' => 1,
        'mappings' => array(
            0 => array(
                'source' => 'xpathparser:0',
                'target' => 'title',
                'unique' => FALSE,
                'language' => 'und',
            ),
            1 => array(
                'source' => 'xpathparser:28',
                'target' => 'field_data_isps_id',
                'unique' => 1,
                'language' => 'und',
            ),
            2 => array(
                'source' => 'xpathparser:1',
                'target' => 'field_data_keywords',
                'unique' => 1,
                'language' => 'und',
            ),
        ),
    ),
),
```
Drupal Integration

Mapping Issues

**RELATED DATABASE FIELD SETTINGS**

These settings apply to the Related database field everywhere it is used. Because the field already has data, some settings can no longer be changed.

**Number of values**

1

Maximum number of values users can enter for this field. "Unlimited" will provide an 'Add more' button so the users can add as many values as they like.

**Maximum length**

255

The maximum length of the field in characters.

**FIELD DATE FIELD SETTINGS**

These settings apply to the Field date field everywhere it is used. Because the field already has data, some settings can no longer be changed.

**Number of values**

1

Maximum number of values users can enter for this field. "Unlimited" will provide an 'Add more' button so the users can add as many values as they like.

**Date attributes to collect**

- Year
- Month
- Day
- Hour
- Minute
- Second

Select the date attributes to collect and store. Changes to date attributes only affects new or updated content.

- Collect an end date

End dates are used to collect duration. E.g., allow an event to start on September 15, and end on September 16.

- Cache dates

Date objects can be created and cached as date fields are loaded, rather than when they are displayed, to improve performance.

Save settings
Drupal Integration

Schedule Importers
YARD Work: Data Archiving and Publishing at Yale

Limor Peer  limor.peer@yale.edu
Joshua Dull  joshua.dull@yale.edu
Themba Flowers  themba.flowers@yale.edu

Acknowledgments: Innovations for Poverty Action, Colectica, Digital Lifecycle Research & Consulting, Yale University Library, YaleSites
Technical Components & Support at Yale

Yale ITS
• Hardware – Windows Server (VM), 32GB RAM minimum (8 Cores), 100GB local disk for OS, applications and swap files
• Software – Colectica repository, statistical software, integrated APIs
• Storage – storage@yale start at 500GB read, write, no-execute access to one or more directories
• Application hosting – WCF application and ASP.NET MVC web application on IIS, plus a SQL Server database (10GB), a Windows Service
• Security – Federated identification

Yale Library
• Persistent links – handle service
• Long-term preservation – Fedora Commons / Hydra*
• Discovery – Blacklight*

*inactive