Crowdsourcing DDI Development: New Features from the CED^2AR Project

Benjamin Perry, Venkata Kambhampaty, Kyle Brumsted, Lars Vilhuber, William Block
What is CED$^2$AR?

- Part of the NSF Census Research Network (NCRN) (Grant #1131848)
- Lightweight, DDI driven web application
- Enables search, browsing and editing across codebooks
- Provides an open API for developers
- Live example at demo.ncrn.cornell.edu
EDDI 2014 “Collaborative editing…”

• Emphasis on collaborative editing (small set of users)
  – Online editor
  – Versioned and tracked metadata through Git
  – Tied into external authentication frameworks
Now

• Support crowdsourced DDI curation through CED$^2$AR
  – Accommodating more users
  – Allow for application specific customization
  – Create incentives and guidance for users
  – Abstract technical barriers
Initial metadata (DDI) has been created and ingested into a CED²AR instance

Metadata may be
- Incomplete (valid DDI but empty or non-informative fields)
- Lacking user feedback (on value or constraints of variables)

Assumption:
- Archivist is not the only specialist on a particular dataset
- Users collectively have information that is not initially included in metadata
User Workflow

1. User searches through CED²AR or external search engine
2. User discovers data relevant to their query
3. User can choose to contribute structured or unstructured documentation for datasets
   - No DDI knowledge required – user documents on fields, without needing to know how that fits into a particular metadata structure
   - May involve creating links (provenance) to other datasets
Attracting Users
1. Search engine optimization enhancements to DDI
2. Exposing community contributions

Retaining Users
1. Flexible authentication
2. Easy to use editor
3. Metadata scoring
4. Tracking and identifying community contributions
Search Engine Optimization

• Expanding the interoperability of DDI

FAQ - CED2AR
https://www2.ncrn.cornell.edu/ced2ar-web/docs/faq
FAQs for the CED2AR project. ... *age matches terms that end with age (wage and marriage would match); *age* matches terms that contain age (wages would ...

Age (IPUMSUSA2012) - CED2AR
dev.ncrn.cornell.edu/ced2ar-web/codebooks/ipumsusa/v/2012/vars/AGE
AGE reports the person’s age in years as of the last birthday. ... CED2AR. The Comprehensive Extensible Data Documentation and Access Repository.

MBR/PHUS Variables Group (SSBV51) - CED2AR
https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssbv/v51/.../_7/ ▼
The Master Benefits Records (MBR) is SSA’s main file to track who is receiving Old Age Survivor and Disability (OASDI) benefits, the reason for receipt, and the ...
Authentication

• **Support OpenID and OAuth2**
  – Currently using Google with OAuth2
  – Developing connectors to work with additional providers

• **CED$^2$AR handles identity management**
Editing

- Automatic validation, and editor for rich content
Editing

• Allows for ASCII Math

The between implicate variance for a generic variable $X$ is:

$$B[\bar{X}_{agt}] = \frac{1}{M-1} \sum_{l=1}^{100} (\text{hat}(X)_{agt}^{(l)} - \bar{X}_{agt})^2$$
Editing

• Growing support for additional DDI fields, exposed or not
Metadata Scoring

- Exposing sparse documentation

---

Codebook Score

Variables

98.4% of variables have labels

Variables without labels
- KPq3_text - RDq2@year_r

0.8% of variables have significant full descriptions

Variables without significant full descriptions

95.1% of variables have values

Variables without values
- CASEID - FNLD - HHSIZE_TOT - KPq3_text - MSA - STATE
## Modified Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Date Changed</th>
<th>Commit Message</th>
<th>User</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>wksjob_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>wksjob_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>wkswp_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>vetrecip_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>vetrecip_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>wcoamt_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
<tr>
<td>vetrecip_MN</td>
<td>March 22, 2015</td>
<td>View commit</td>
<td><a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a></td>
<td>Remote Change</td>
</tr>
</tbody>
</table>
Versioning

- Uses Git, a distributed version control system
- Every aspect of the system is configurable
  - Scheduled tasks check for changes
  - Once changes exceed threshold, they are pushed
  - Pending changes are pushed after a time limit or on demand
Architecture

Master Branch (Official version)

1. User gets copy of DDI to edit

User Contributed Branch
Architecture

1. User gets copy of DDI to edit

2. Each edit is versioned
1. User gets copy of DDI to edit

2. Each edit is versioned

3. Data provider merges user’s edits back into official DDI

Master Branch (Official version)
# Architecture

## Branches

<table>
<thead>
<tr>
<th>Branch</th>
<th>Behind</th>
<th>Ahead</th>
<th>Updated</th>
<th>Pull request</th>
</tr>
</thead>
<tbody>
<tr>
<td>master</td>
<td></td>
<td></td>
<td>2014-11-13</td>
<td></td>
</tr>
<tr>
<td>venkytest</td>
<td>75</td>
<td></td>
<td>43 seconds ago</td>
<td></td>
</tr>
<tr>
<td>benlocal</td>
<td>46</td>
<td></td>
<td>21 hours ago</td>
<td></td>
</tr>
<tr>
<td>ssbtesting</td>
<td>33</td>
<td></td>
<td>4 days ago</td>
<td></td>
</tr>
<tr>
<td>localssb</td>
<td>58</td>
<td></td>
<td>2015-03-10</td>
<td></td>
</tr>
<tr>
<td>acsdev</td>
<td>12</td>
<td></td>
<td>2015-02-27</td>
<td></td>
</tr>
<tr>
<td>acsdev_test</td>
<td>2</td>
<td></td>
<td>2015-02-25</td>
<td></td>
</tr>
<tr>
<td>testing</td>
<td>4</td>
<td></td>
<td>2015-02-18</td>
<td></td>
</tr>
<tr>
<td>cestesting</td>
<td>6</td>
<td></td>
<td>2015-01-28</td>
<td></td>
</tr>
</tbody>
</table>
Architecture

Database  \(\leftrightarrow\) Web Application  \(\leftrightarrow\) Local Repository

Server

Remote Repository
Architecture

CED$^2$AR Instance

Remote Repository
Architecture

CED²AR Instance

CED²AR Instance

CED²AR Instance

CED²AR Instance

Remote Repository
Architecture

CED²AR Instance

CED²AR Instance

CED²AR Instance

CED²AR Instance

Remote Repository

CED²AR Instance (Official)
Remote Location

- Our implementation uses Bitbucket
- Commit messages describe changes
- Users linked by email address
- Commit hashes are stored on CED^2AR
- Remote synchronization is optional
Remote Location

Anonymous committed 6f9d8f6

{ssbv51,bap63@cornell.edu,var,birthdate}
{ssbv51,bap63@cornell.edu,cover}

diamond 3bf58ce
• cestesting
 View raw commit
 Watch this commit
## Tracking Changes

### Codebook Status

<table>
<thead>
<tr>
<th>Codebook</th>
<th>Git Status</th>
<th>Last Local Update</th>
<th>BaseX Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>acs.2009.xml</td>
<td>UP_TO_DATE</td>
<td>February 25, 2015 at 11:05 AM: Committing codebooks retrieved directly from BaseX</td>
<td>DOES_NOT_EXIST_IN_BASEX</td>
</tr>
<tr>
<td>acs.2012-dw.xml</td>
<td>UP_TO_DATE</td>
<td>March 25, 2015 at 11:09 AM: Auto commit on application shutdown</td>
<td>DOES_NOT_EXIST_IN_BASEX</td>
</tr>
<tr>
<td>acs.2012.xml</td>
<td>UP_TO_DATE</td>
<td>March 25, 2015 at 11:17 AM: Committing codebooks retrieved directly from BaseX</td>
<td>EXIST_IN_BASEX</td>
</tr>
<tr>
<td>cnss.2012.xml</td>
<td>UP_TO_DATE</td>
<td>March 25, 2015 at 11:11 AM:</td>
<td>EXIST_IN_BASEX</td>
</tr>
<tr>
<td>ecf.1.xml</td>
<td>UP_TO_DATE</td>
<td>March 12, 2015 at 9:36 AM: Committing codebooks retrieved directly from BaseX</td>
<td>EXIST_IN_BASEX</td>
</tr>
<tr>
<td>heg.3.xml</td>
<td>UP_TO_DATE</td>
<td>March 25, 2015 at 12:28 PM:</td>
<td>EXIST_IN_BASEX</td>
</tr>
<tr>
<td>ipumsusa.2012.xml</td>
<td>UP_TO_DATE</td>
<td>March 25, 2015 at 12:46 PM:</td>
<td>EXIST_IN_BASEX</td>
</tr>
</tbody>
</table>


Continued Work: Improving merge control

1. User gets copy of DDI to edit

2. Each edit is versioned

3. Data provider merges user’s edits back into official DDI
Continued Work: The uncontrolled merge

- Workflow as described assumes metadata curator merges information
- Within the limits of a 24-hour day: what’s the likelihood that that process scales?
- Alternate: “wiki” methodology
Architecture (alternate)

Master Branch
(Official version)

Wiki Branch
(Community version)

Codebook 1.0
Architecture (alternate)

Master Branch (Official version)

Codebook 1.0

Wiki Branch (Community version)

Codebook 1.0

User Branches

1. Users pull from wiki branch into any instance of CED²AR

2.
Architecture (alternate)

Master Branch (Official version)

Codebook 1.0

Wiki Branch (Community version)

Codebook 1.0

Codebook rev 1

User Branches

Users push back to branch manually
Architecture (alternate)

Master Branch
(Official version)

Codebook 1.0

Wiki Branch
(Community version)

Codebook 1.0

Codebook rev 1

User Branches

New users work off most recent revision by default
Architecture (alternate)

Master Branch
(Official version)

Codebook 1.0

Wiki Branch
(Community version)

Codebook 1.0
Codebook rev 1

User Branches

Codebook rev X
Architecture (alternate)

Master Branch (Official version)

Wiki Branch (Community version)

Codebook 1.0

Codebook rev 1

Codebook rev X

User Branches
Architecture (alternate)

Master Branch
(Official version)

Codebook 1.0

Wiki Branch
(Community version)

Codebook 1.0

User Branches

Codebook rev 1

User is responsible for merging

Codebook rev X
Architecture (alternate)

Master Branch
(Official version)
Codebook
1.0

Wiki Branch
(Crowdsourced version)
Codebook
rev X

CED²AR User
Interface exposes both
versions
(with attribution)
Continued Work: Improving merge control

- Merging crowd-sourced content back into official documentation
Thank you!
Questions?
ced2ar-devs-l@cornell.edu
ncrn.cornell.edu
github.com/ncrnncornell
Extra slides
Continued Work: Facilitating Editing

- Tagging variables with a controlled vocabulary and a folksonomy

The between implicate variance for a generic variable $X$ is:

$$B[X_{aikt}] = \frac{1}{M-1} \sum_{t=1}^{100} (\bar{X}_{aikt} - \bar{X}_{aikt})^2$$
Ingest Workflow

SAS
STATA
SPSS
R

Clean, normalize and generate metadata

SQL Metadata

Generate XML

XSD Schema

CED²AR

Data/Misc Files