Collaborative Editing of DDI Metadata: The Latest from the CED²AR Project

Benjamin Perry, Venkata Kambhampaty, Kyle Brumsted, Lars Vilhuber, William C. Block
Outline

• Introduction
• Problem
• Solution
• Implementation
• Future Work
What is CED$^2$AR?

- Funded by NSF grant #1131848
- Lightweight, DDI driven web application
- Designed specifically for custom DDI schema
- Enables search and browsing across codebooks
- Provides an open API for developers
- Online at www2.ncrn.cornell.edu/ced2ar-web
The Challenge

- Infrastructure disconnected from curation process
- Researchers not familiar with XML and DDI
- Metadata was not being preserved
The Solution

• Build from existing application
• Keep lightweight infrastructure
• Automate as much as possible
• Prevent steep learning curve
Current Architecture

Server Instance

- Tomcat 7 Server
  - Web Application
  - Database

Desktop Instance

- JAR file
  - Web Application
  - Database
Process

1. User logs into CED²AR
2. User uploads sparse DDI
3. CED²AR validates and cleans DDI
4. Users edit codebooks
5. Git passively versions edits
6. Changes are pushed to remote location
Structure

Web Application + Editing Forms

Database

Local Repo

Remote Repo

3rd Party Authentication

or

Internal Authentication
1. Authentication

- Support OpenID and OAuth2
  - Currently using Google with OAuth2
- CED$^2$AR handles identity management
2. Uploading and Ingest

- Validates against DDI 2.5 schema
- Inserts dates, citations, software references, etc.
- Indexes and assigns codebook internal handle
2. Uploading and Ingest
3. Editing Process

- User searches or browses
- Web forms provide control over content
- CED^2AR supports basic HTML and ASCII math
4. Editing Process

• Built in documentation
4. Editing Process

• Editing a variable

Full Description

ABSENT indicates whether persons who did not work during the previous week had a job or business from which they were temporarily absent and, if so, whether they were absent due to a layoff or if their absence was for some other reason. Persons who responded "no" to the question, "Was this person temporarily absent or on layoff from a job or business last week?" would be considered either unemployed or not in the labor force, depending upon their responses to other questions. See EMPSTAT for definitions of key labor force and employment terminology.

Files

ExtractData: usa_00607.dat (No hyperlink available) (ISO-8859-1 data file)

Values

<table>
<thead>
<tr>
<th>Value</th>
<th>Category</th>
<th>Category Values</th>
<th>Access Level</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N/A</td>
<td>Access Level: undefined</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>No</td>
<td>Access Level: undefined</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Yes, laid off</td>
<td>Access Level: undefined</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Yes, other reason (vacation, illness, labor dispute, etc.)</td>
<td>Access Level: undefined</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Not reported</td>
<td>Access Level: undefined</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Notes
4. Editing Process

• Control over multiple access levels
5. Versioning

- Uses Git, a distributed version control system
- Scheduled tasks check for changes
- Once changes exceed threshold, they are pushed
- Pending changes are pushed after a time limit
6. Remote Location

- Our implementation uses Bitbucket
- Commit messages describe changes
- Users linked by email address
- Commit hashes are stored on CED²AR
6. Versioning

• Viewing version history
6. Remote Location

Anonymous committed 6f9d8f6 yesterday
{ssbv51,bap63@cornell.edu,var.birthdate}
{ssbv51,bap63@cornell.edu,cover}

3bf58ce
cesting
View raw commit
Watch this commit

ssb.v51.xml

11 11  <AuthEnity affiliation="Cornell University">.Virtual RDC</AuthEnity>
12 12  </rspStmt>
13 13  </proStmnt>
14 14  <producer abbr="Cornell NCRN Project">Cornell NSF-Census Research Network (NCRN)</producer>
15 15  + <producer abbr="Cornell NCRN Project">Cornell NSF Census Research Network</producer>
16 16  <copyright>Cornell NCRN Project</copyright>
17 17  + <copyright>Cornell NCRN Project</copyright>
18 18  <prodDate date="2014-06-18">June 18th 2014</prodDate>
19 19  + <prodDate date="2014-06-18">June 18th 2014</prodDate>
20 20  <prodPlac>Cornell Institute for Social and Economic Research (CISER), Cornell University, Ithaca NY</prodPlac>
21 21  + <prodPlac>Cornell Institute for Social and Economic Research (CISER), Cornell University, Ithaca NY</prodPlac>
22 22  <software>CEODAR, Version 1.0.0/</software>
23 23  + <software>The Comprehensive Extensible Data Documentation and Access Repository 2.5</software>
24 24  <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
25 25  + <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
26 26  <grantNo agency="National Science Foundation">1131848</grantNo>
27 27  + <grantNo agency="National Science Foundation">1131848</grantNo>
28 28  <prodStmnt>
29 29  + <prodStmnt>
30 30  <version date="2014-10-07 09:10:48 (auto-generated)">2014-06-18</version>
31 31  + <version date="2014-10-07 09:10:48 (auto-generated)">2014-06-18</version>
32 32  <verStnt>
33 33  + <verStnt>
34 34  <bibtCit>Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook</bibtCit>
35 35  + <bibtCit>Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook
Use Case

CED²AR Server Instance

Remote Repo

CED²AR Desktop Instance

CED²AR Desktop Instance
Future Work

- Focusing on crowdsourcing
- Enhancements to our DDI editor
- Open a demo up to the public
Thank you!

Questions?

ced2ar-devs-l@cornell.edu