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

Benjamin Perry, Venkata Kambhamaty, 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](http://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 CED\textsuperscript{2}AR
2. User uploads sparse DDI
3. CED\textsuperscript{2}AR 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

Modify a Codebook
3. Editing Process

- User searches or browses
- Web forms provide control over content
- CED²AR supports basic HTML and ASCII math
4. Editing Process

- Built in documentation
4. Editing Process

• Editing a variable
4. Editing Process

- Control over multiple access levels

![Variable Access - SIPP Synthetic Beta v51](image)
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

SIPP Synthetic Beta v5.1

View Variables (102 variables)
Last update to metadata: 2014-11-13 10:38:45 (auto-generated)
Document Date: June 19th 2014

Codebook prepared by: Cornell NSF Census Research Network

Data prepared by: United States Department of Commerce, Bureau of the Census.
6. Remote Location

- Our implementation uses Bitbucket
- Commit messages describe changes
- Users linked by email address
- Commit hashes are stored on CED$^2$AR
6. Versioning

• Viewing version history
6. Remote Location

Anonymous committed 6f9d8f6

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

◊ 3bf58ce
◦ cestesting
◦ View raw commit
◦ Watch this commit
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