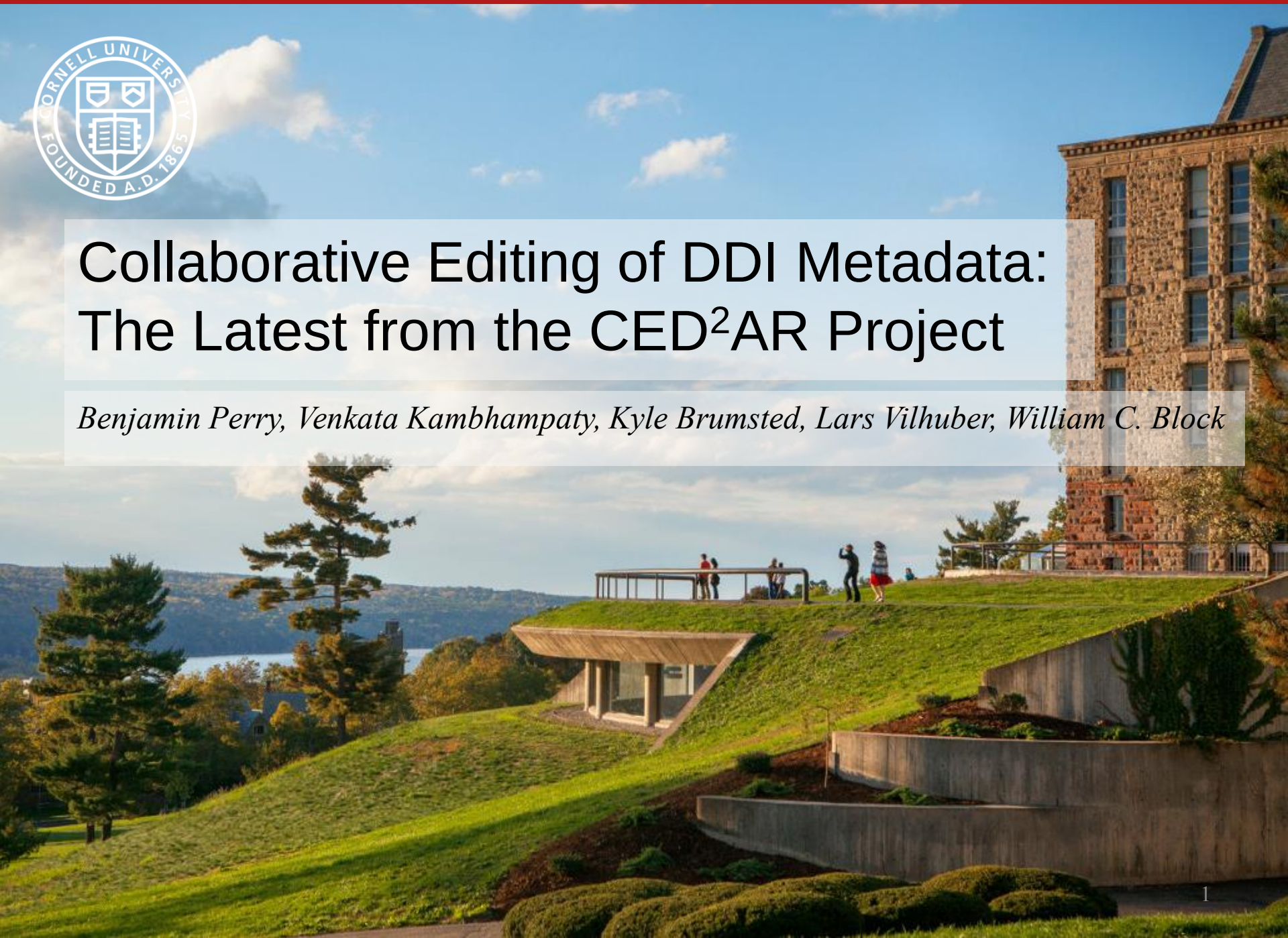


Collaborative Editing of DDI Metadata: The Latest from the CED²AR Project

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

Outline

- Introduction
- Problem
- Solution
- Implementation
- Future Work

What is CED²AR?

- 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

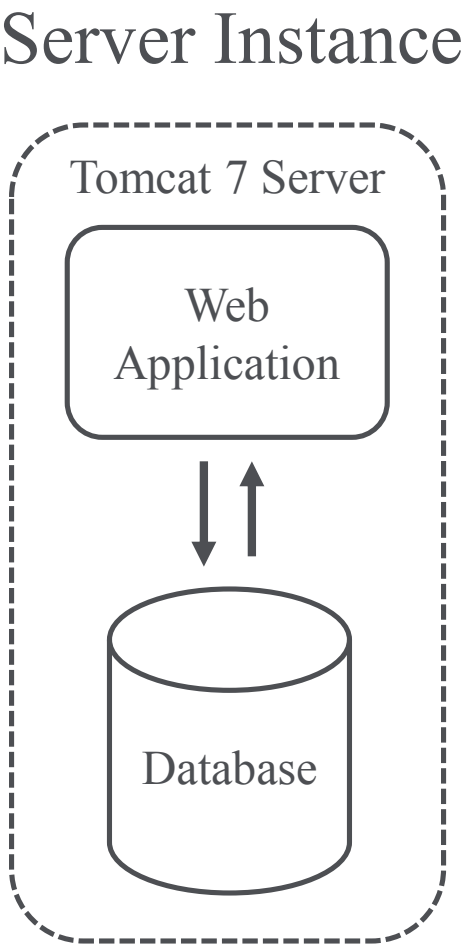
The Challenge

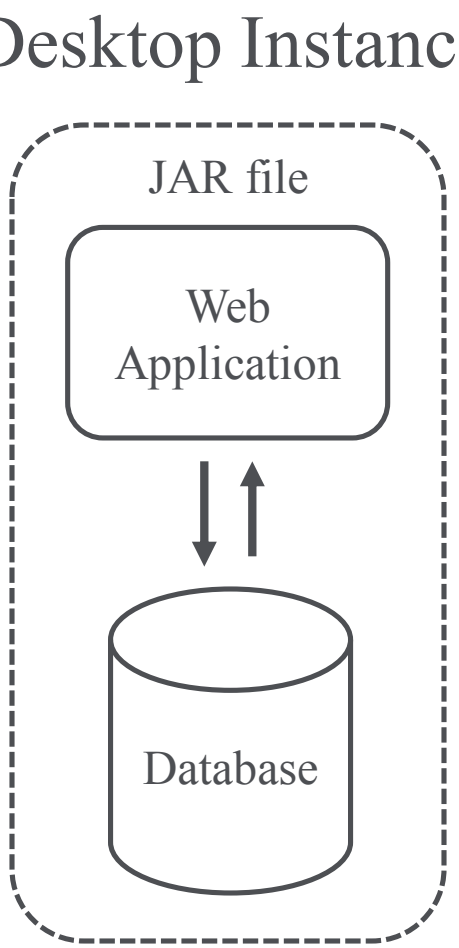
- Infrastructure disconnected from curation process
- Researchers not familiar with XML and DDI
- Metadata was not being preserved

The Solution

- Build from existing application
- Keep lightweight infrastructure
- Automate as much as possible
- Prevent steep learning curve

Current Architecture

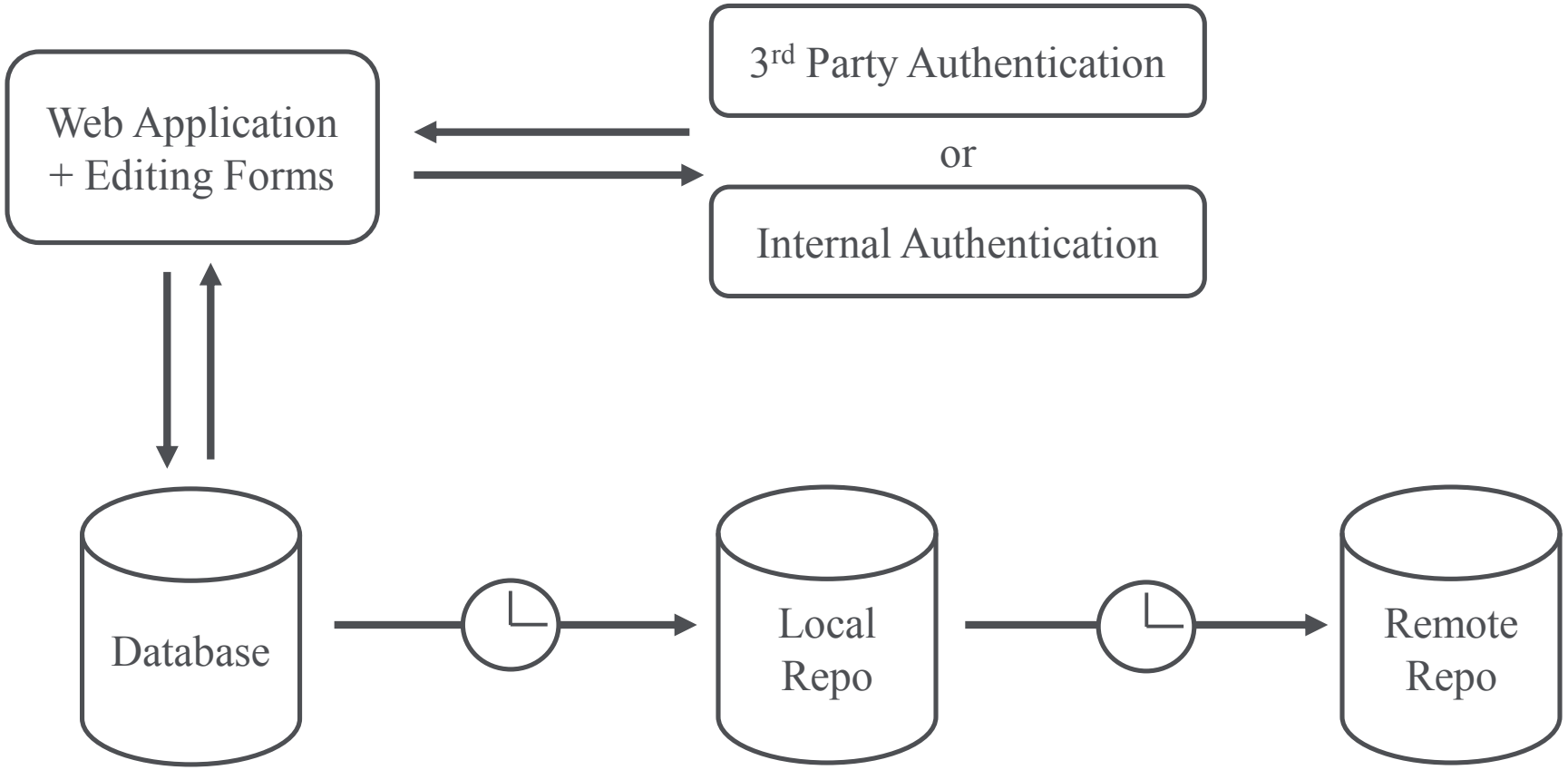




Process

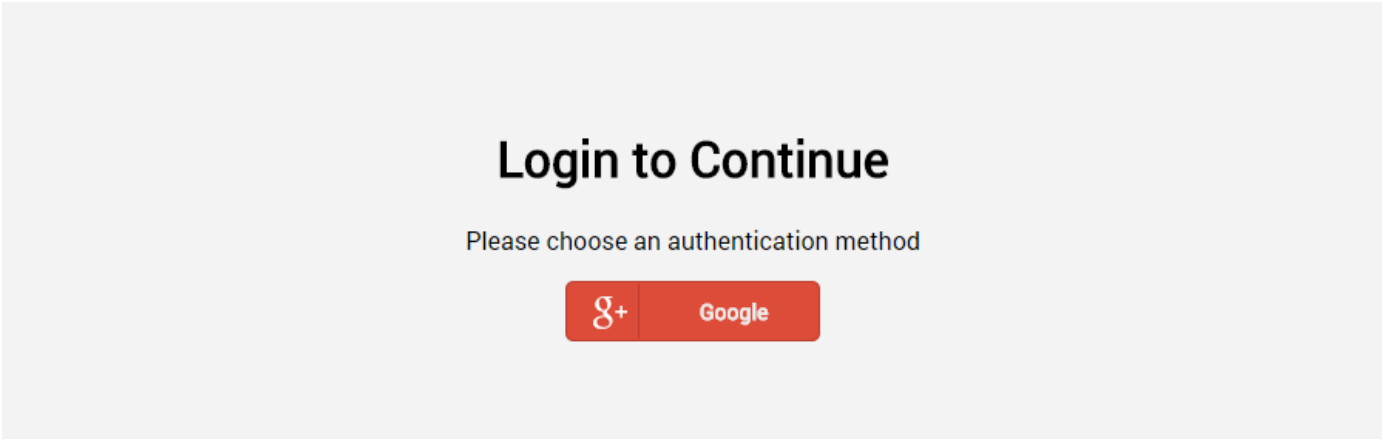
1. User logs into CED^{2}AR
2. User uploads sparse DDI
3. CED^{2}AR validates and cleans DDI
4. Users edit codebooks
5. Git passively versions edits
6. Changes are pushed to remote location

Structure


Web Application
+ Editing Forms
3rd Party Authentication
or
Internal Authentication
Database
Local
Repo
Remote
Repo

1. Authentication

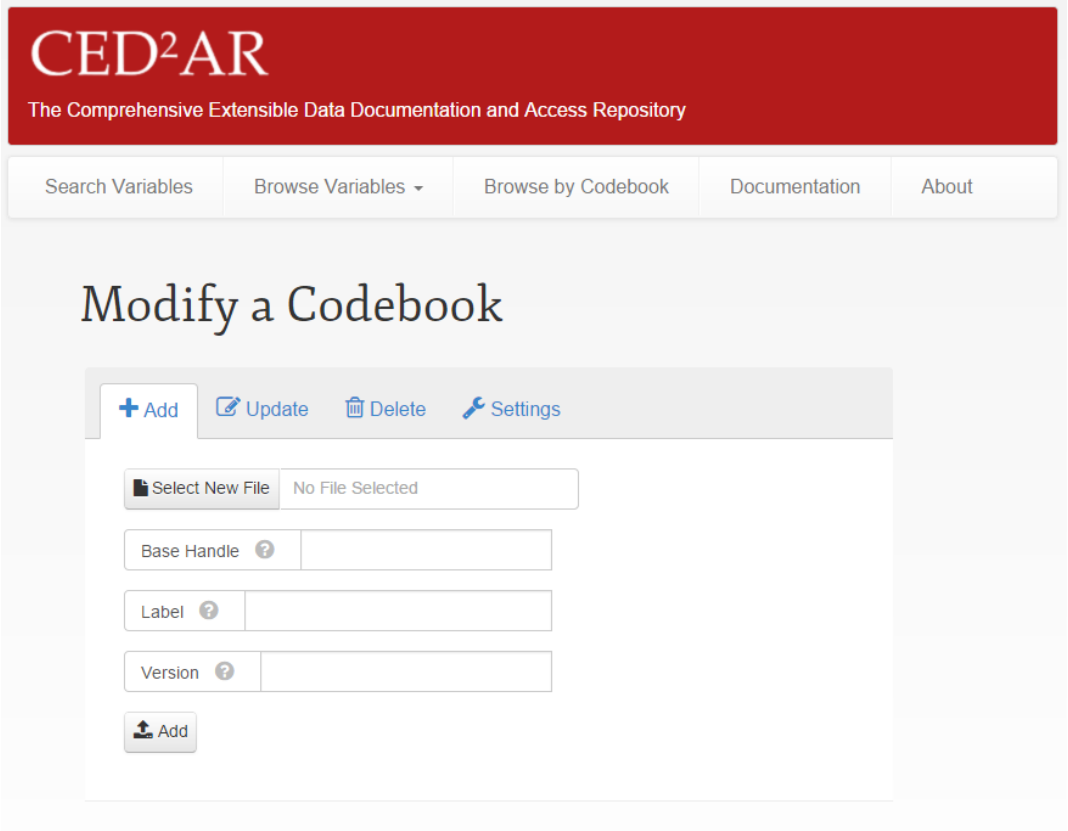
- Support OpenID and OAuth2
  - Currently using Google with OAuth2
- CED²AR handles identity management



2. Uploading and Ingest

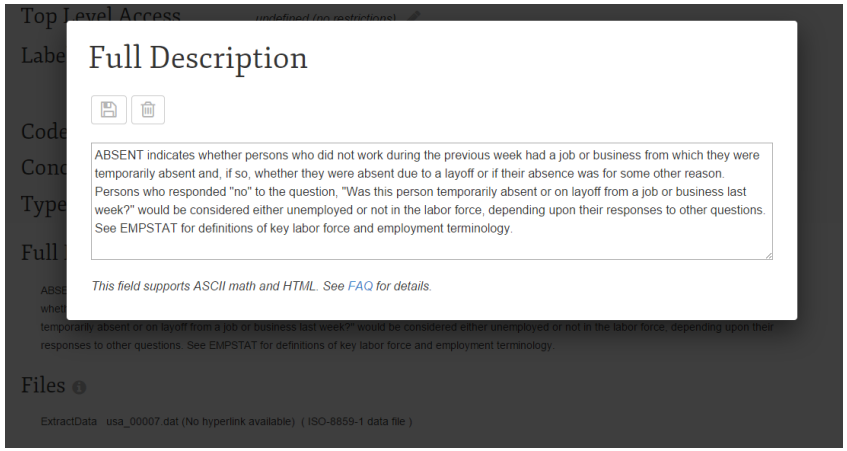
- Validates against DDI 2.5 schema
- Inserts dates, citations, software references, etc.
- Indexes and assigns codebook internal handle

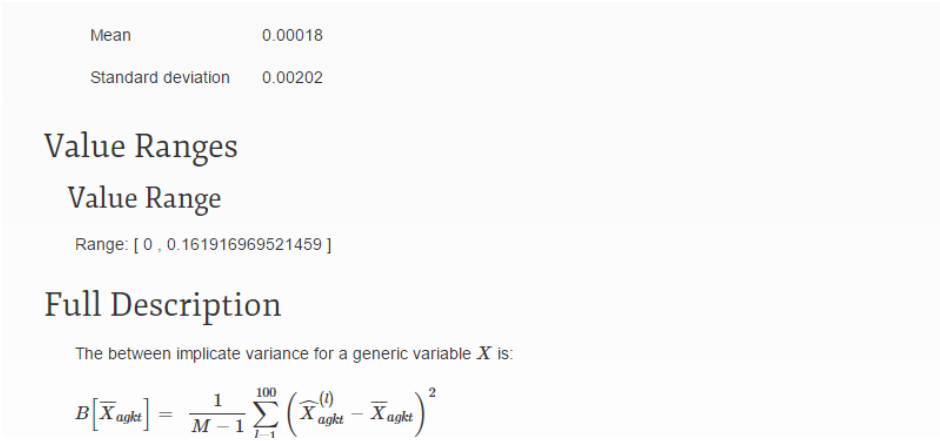
2. Uploading and Ingest



3. Editing Process

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



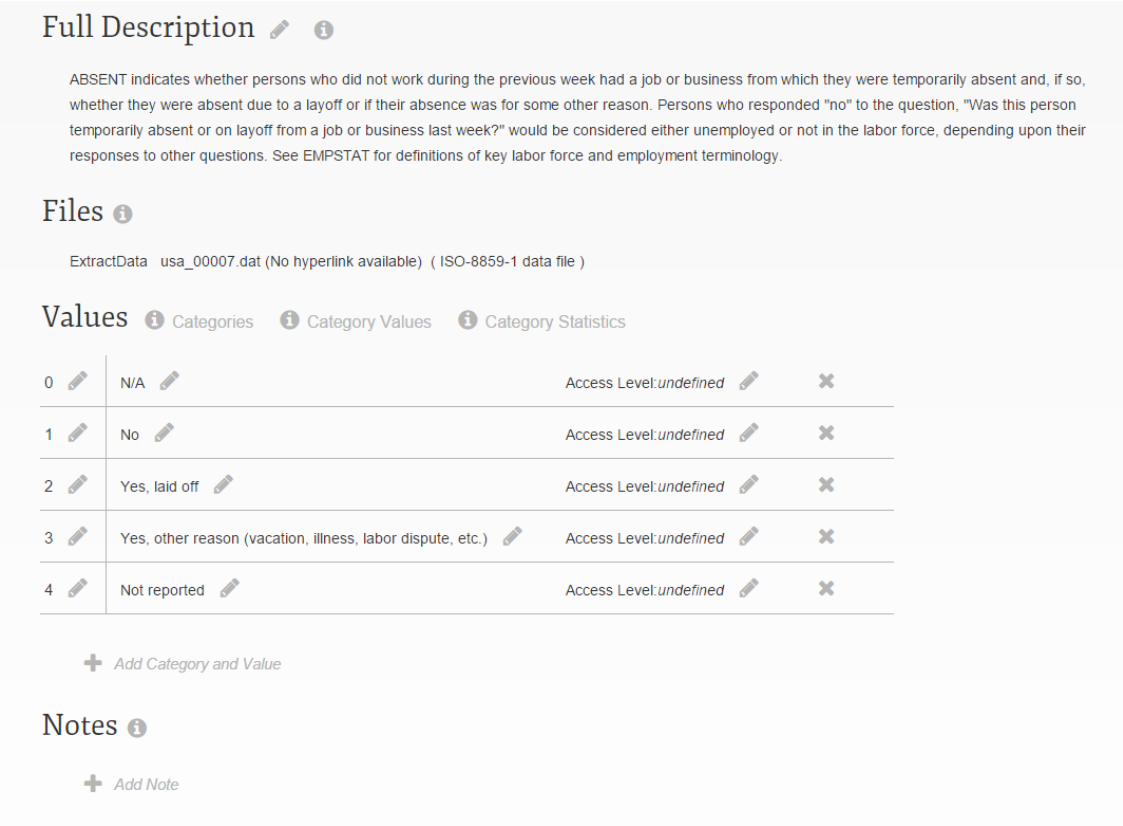


4. Editing Process

- 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_00007.dat (No hyperlink available) (ISO-8859-1 data file)

Values Categories Category Values Category Statistics

0	N/A	Access Level:*undefined*
1	No	Access Level:*undefined*
2	Yes, laid off	Access Level:*undefined*
3	Yes, other reason (vacation, illness, labor dispute, etc.)	Access Level:*undefined*
4	Not reported	Access Level:*undefined*

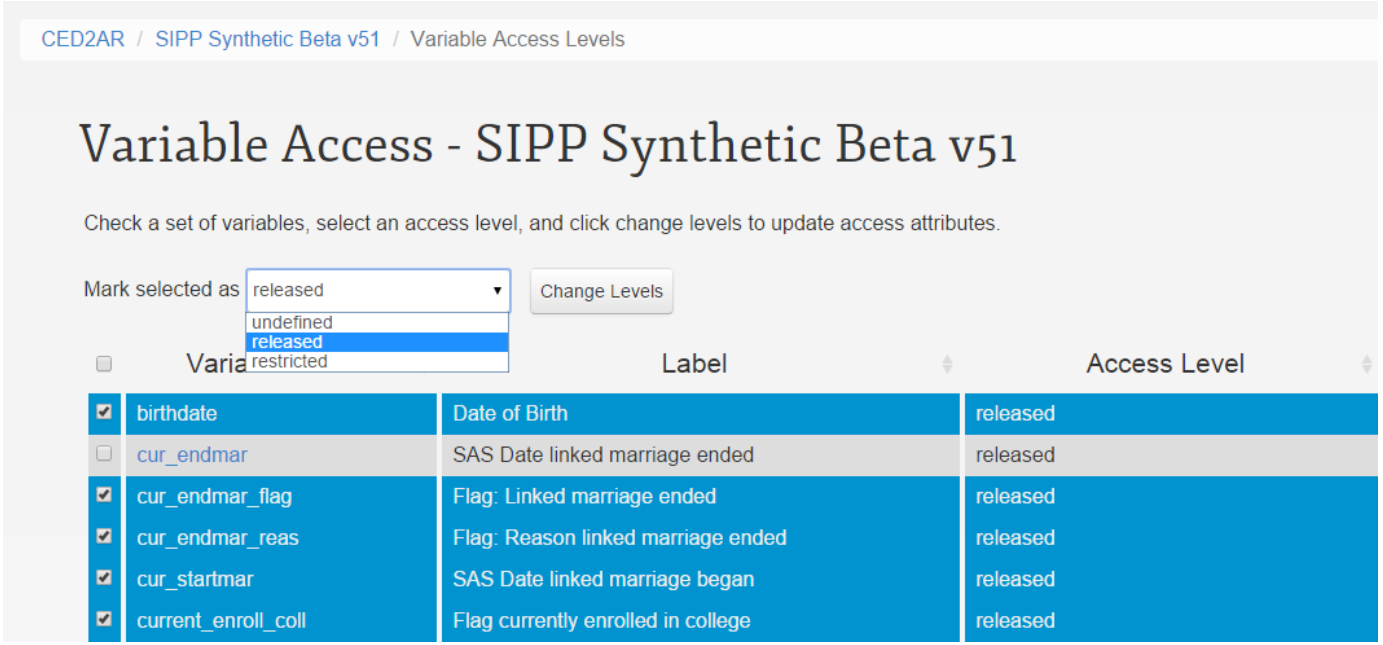
Add Category and Value

Notes

Add Note

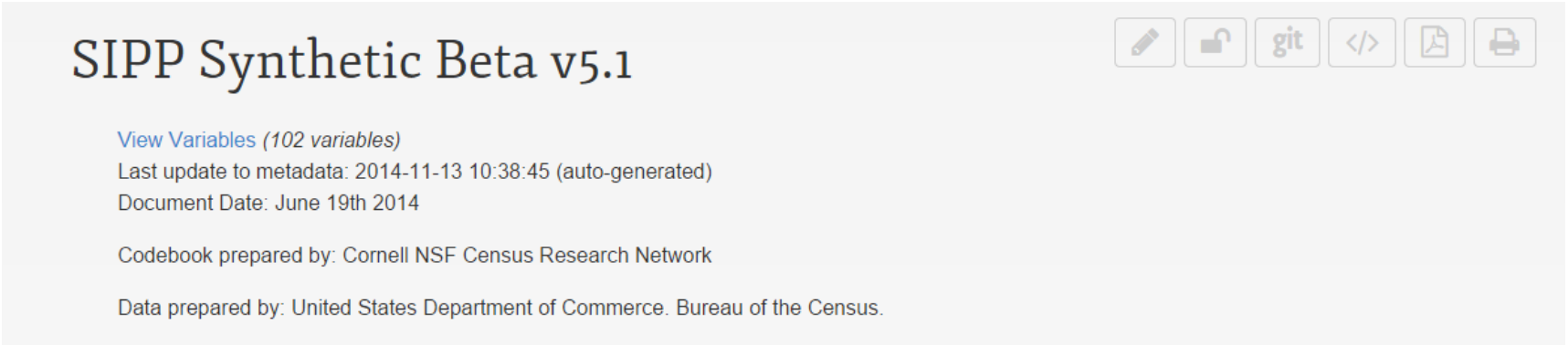
4. Editing Process

- 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

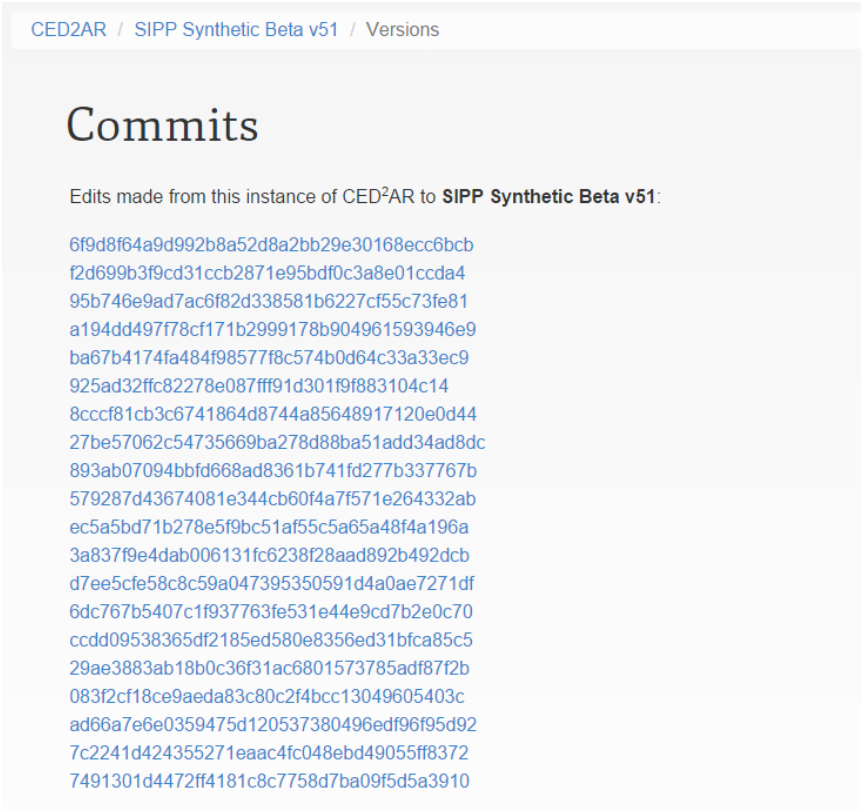


6. Remote Location

- Our implementation uses Bitbucket
- Commit messages describe changes
- Users linked by email address
- Commit hashes are stored on CED^{2}AR

6. Versioning

- Viewing version history



6. Remote Location

Anonymous committed `6f9d8f6`

yesterday

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

`3bf58ce`

cestesting

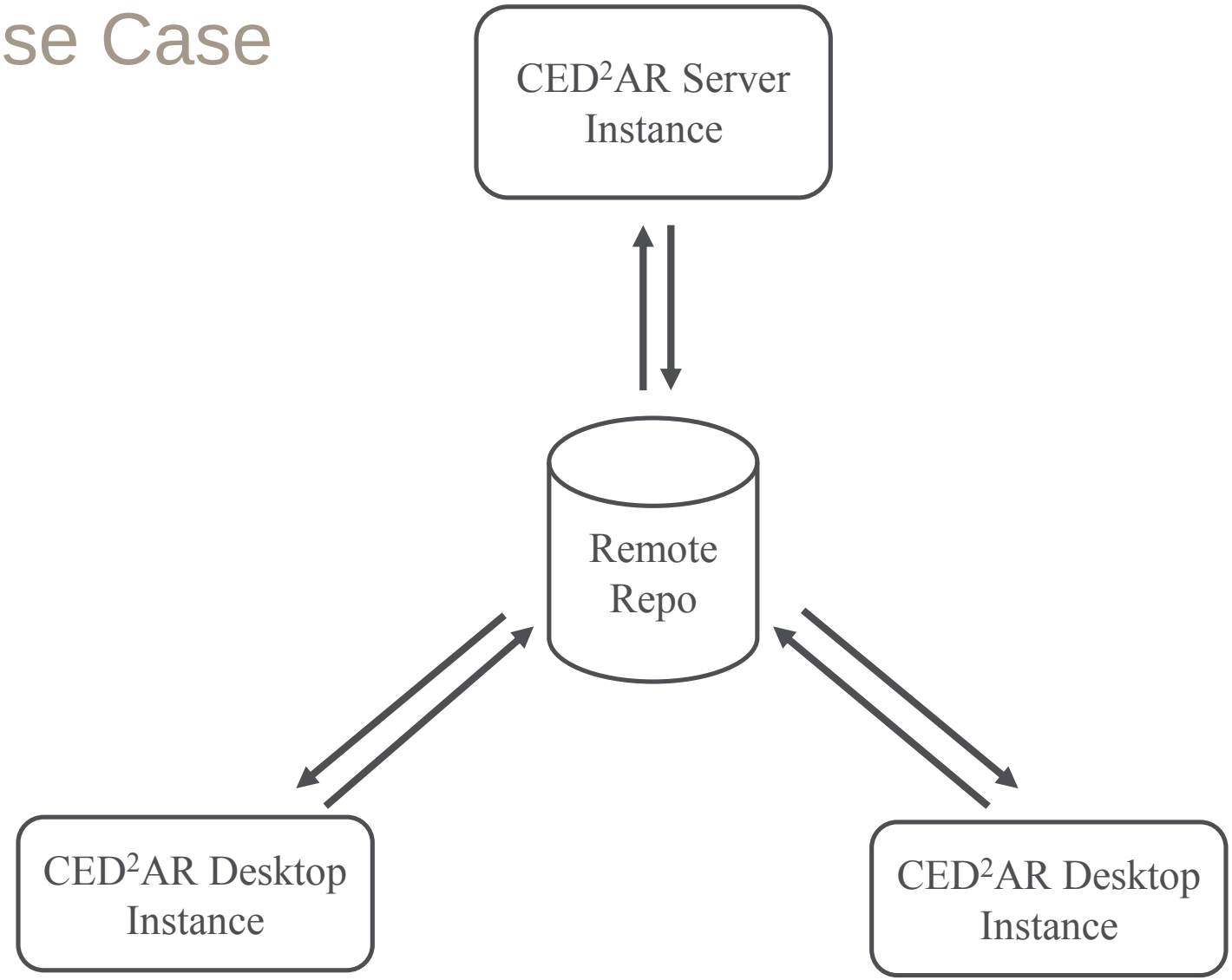
View raw commit

Watch this commit

ssb.v51.xml

```
...
11  11        <AuthEnty affiliation="Cornell University">Virtual RDC</AuthEnty>
12  12      </rspStmt>
13  13      <prodStmt>
14      -     <producer abbr="Cornell NCRN Project">Cornell NSF-Census Research Network (NCRN)</producer>
    14  +     <producer abbr="Cornell NCRN Project">Cornell NSF Census Research Network</producer>
15  15        <copyright>Cornell NCRN Project</copyright>
16      -     <prodDate date="18 June 2014">June 21, 2014</prodDate>
    16  +     <prodDate date="18 June 2014">June 19th 2014</prodDate>
17  17        <prodPlac>Cornell Institute for Social and Economic Research (CISER), Cornell University, Ithaca NY<ExtLink URI="http://ciser
18  18        </prodPlac>
19  19        <software>CED2AR, Version 1.0</software>
    20  +     <software>The Comprehensive Extensible Data Documentation and Access Repository 2.5</software>
20  21        <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
21  22        <grantNo agency="National Science Foundation">1131848</grantNo>
22  23      </prodStmt>
...
27  28        <distDate date="2014">2014</distDate>
28  29      </distStmt>
29  30      <verStmt>
30      -     <version date="2014-10-07 09:10:40 (auto-generated)">2014-06-18</version>
    31  +     <version date="2014-11-13 10:38:45 (auto-generated)">2014-06-18</version>
31  32      </verStmt>
32  33      <biblCit>Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook
33  34    </citation>
```

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