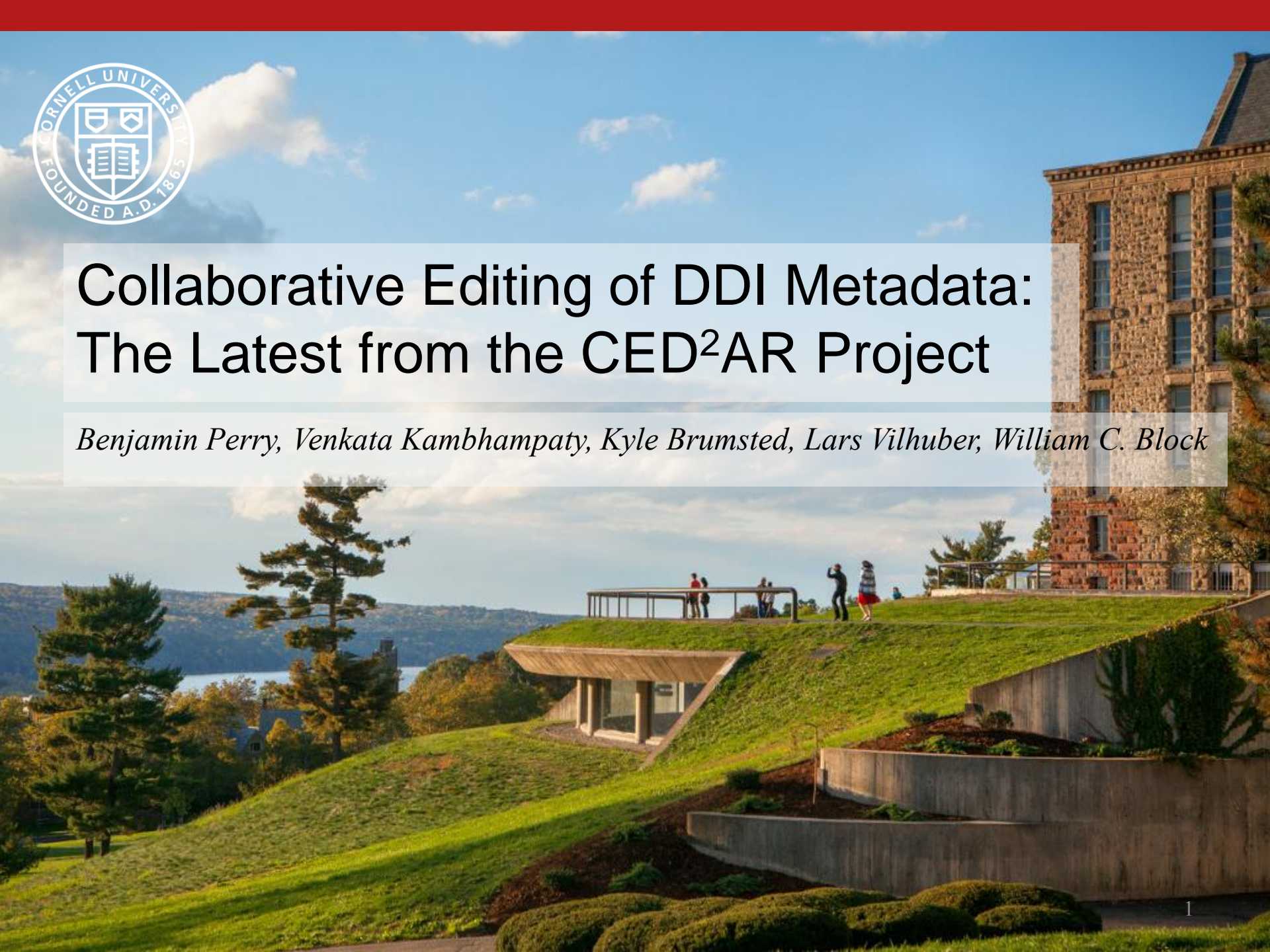




Collaborative Editing of DDI Metadata: The Latest from the CED²AR Project

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



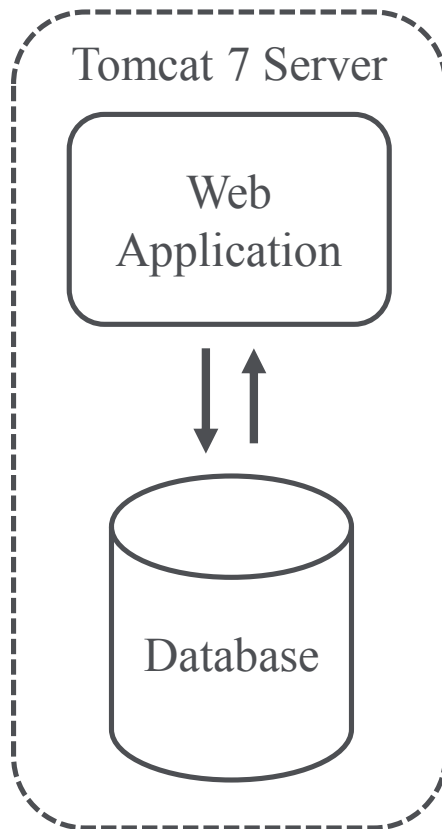
The Solution

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

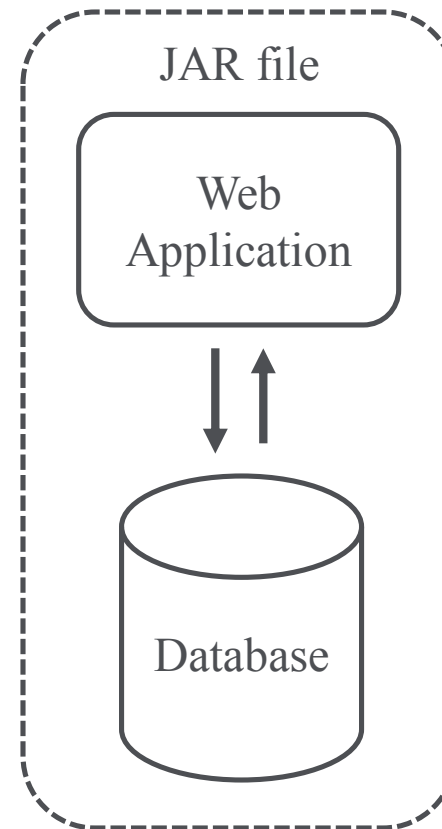


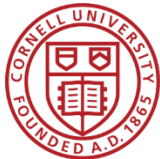
Current Architecture

Server Instance



Desktop Instance



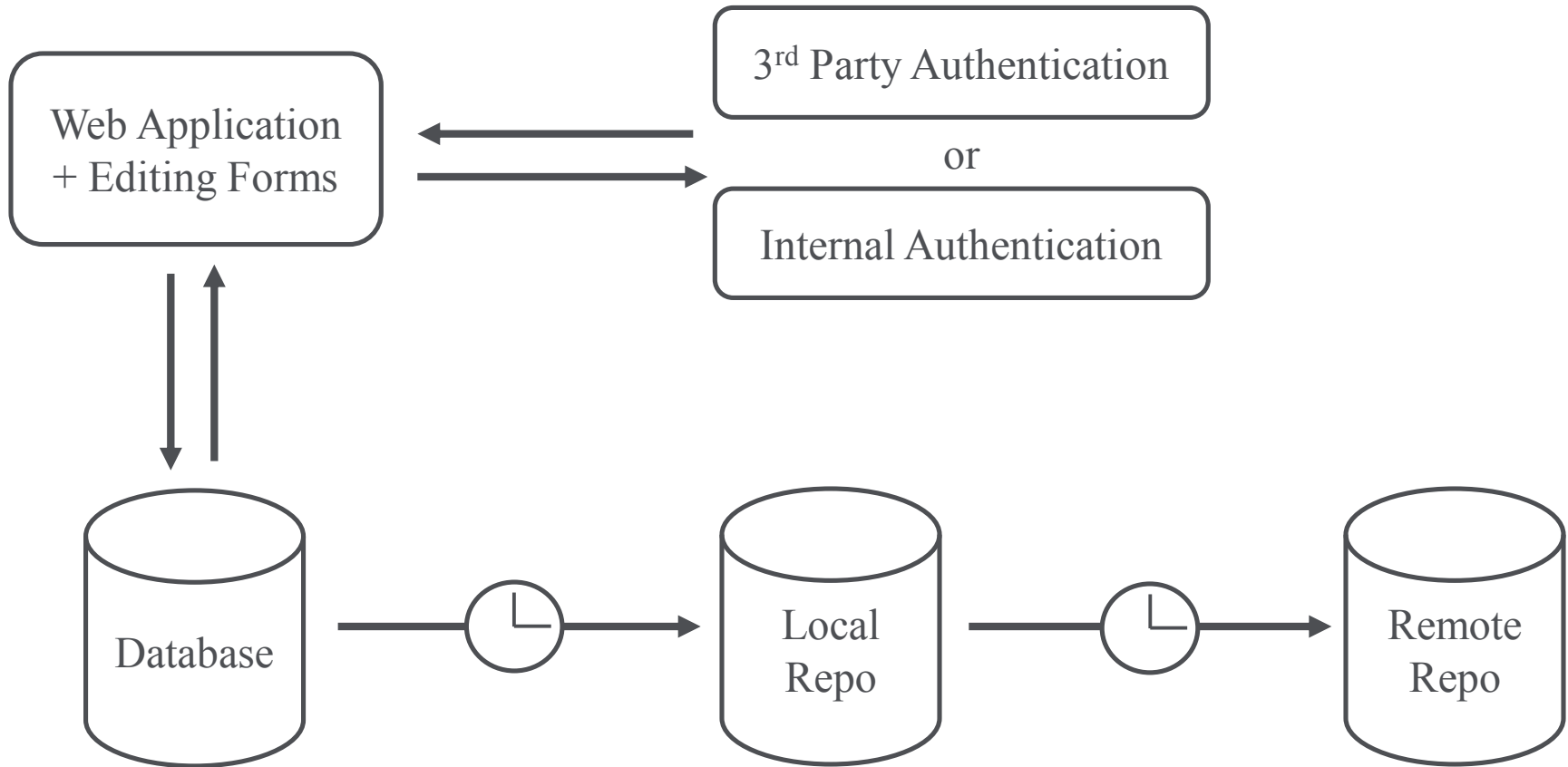


Process

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





1. Authentication

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

Login to Continue

Please choose an authentication method





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

CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

[Search Variables](#) [Browse Variables ▾](#) [Browse by Codebook](#) [Documentation](#) [About](#)

Modify a Codebook

[+ Add](#) [✎ Update](#) [🗑 Delete](#) [🔧 Settings](#)

[📁 Select New File](#) No File Selected

Base Handle [?](#)

Label [?](#)

Version [?](#)

[📁 Add](#)



3. Editing Process

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

Top Level Access

Label

Full Description

Code

Content

Type

Full

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.

This field supports ASCII math and HTML. See [FAQ](#) for details.

Files

ExtractData usa_00007.dat (No hyperlink available) (ISO-8859-1 data file)

Mean 0.00018

Standard deviation 0.00202

Value Ranges

Value Range

Range: [0 , 0.161916969521459]

Full Description

The between implicate variance for a generic variable X is:

$$B[\bar{X}_{agkt}] = \frac{1}{M-1} \sum_{i=1}^{100} \left(\hat{X}_{agkt}^{(i)} - \bar{X}_{agkt} \right)^2$$



4. Editing Process

- Built in documentation

Last update to metadata: 2014-11-13 10:30:43 (auto-generated)

Document Date: June 19th 2014

Bibliographic Citation

Description

Complete bibliographic reference containing all of the standard elements of a citation that can be used to cite the work. The "format" attribute is provided to enable specification of the particular citation style used, e.g., APA, MLA, Chicago, etc.

Example

```
<biblCit format="MRDF">Rabier, Jacques-Rene, and Ronald Inglehart. EURO-BAROMETER 11: YEAR OF THE CHILD IN EUROPE, APRIL 1979 [Codebook file]. Conducted by Institut Francais D'Opinion Publique (IFOP), Paris, et al. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 1981.</biblCit>
```

Citation



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: <i>undefined</i> 	
1 	No 	Access Level: <i>undefined</i> 	
2 	Yes, laid off 	Access Level: <i>undefined</i> 	
3 	Yes, other reason (vacation, illness, labor dispute, etc.) 	Access Level: <i>undefined</i> 	
4 	Not reported 	Access Level: <i>undefined</i> 	

 Add Category and Value

Notes

 Add Note



4. Editing Process

- Control over multiple access levels

CED2AR / SIPP Synthetic Beta v51 / Variable Access Levels

Variable Access - SIPP Synthetic Beta v51

Check a set of variables, select an access level, and click change levels to update access attributes.

Mark selected as released Change Levels

☐ undefined
☒ released
☐ restricted

<input type="checkbox"/>	Variable	Label	Access Level
<input checked="" type="checkbox"/>	birthdate	Date of Birth	released
<input type="checkbox"/>	cur_endmar	SAS Date linked marriage ended	released
<input checked="" type="checkbox"/>	cur_endmar_flag	Flag: Linked marriage ended	released
<input checked="" type="checkbox"/>	cur_endmar_reas	Flag: Reason linked marriage ended	released
<input checked="" type="checkbox"/>	cur_startmar	SAS Date linked marriage began	released
<input checked="" type="checkbox"/>	current_enroll_coll	Flag currently enrolled in college	released



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



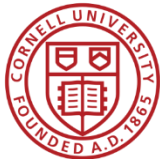
[View Variables](#) (102 variables)

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 CED²AR



6. Versioning

- Viewing version history

[CED2AR](#) / [SIPP Synthetic Beta v51](#) / [Versions](#)

Commits

Edits made from this instance of CED²AR to **SIPP Synthetic Beta v51**:

6f9d8f64a9d992b8a52d8a2bb29e30168ecc6bcb
 f2d699b3f9cd31ccb2871e95bdf0c3a8e01ccda4
 95b746e9ad7ac6f82d338581b6227cf55c73fe81
 a194dd497f78cf171b2999178b904961593946e9
 ba67b4174fa484f98577f8c574b0d64c33a33ec9
 925ad32ffc82278e087fff91d301f9f883104c14
 8ccc81cb3c6741864d8744a85648917120e0d44
 27be57062c54735669ba278d88ba51add34ad8dc
 893ab07094bbfd668ad8361b741fd277b337767b
 579287d43674081e344cb60f4a7f571e264332ab
 ec5a5bd71b278e5f9bc51af55c5a65a48f4a196a
 3a837f9e4dab006131fc6238f28aad892b492dcb
 d7ee5cfe58c8c59a047395350591d4a0ae7271df
 6dc767b5407c1f937763fe531e44e9cd7b2e0c70
 ccdd09538365df2185ed580e8356ed31bfca85c5
 29ae3883ab18b0c36f31ac6801573785adf87f2b
 083f2cf18ce9aada83c80c2f4bcc13049605403c
 ad66a7e6e0359475d120537380496edf96f95d92
 7c2241d424355271eaac4fc048ebd49055ff8372
 7491301d4472ff4181c8c7758d7ba09f5d5a3910



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 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      - <prodDate date="18 June 2014">June 21, 2014</prodDate>
18 18      + <prodDate date="18 June 2014">June 19th 2014</prodDate>
19 19      <prodPlac>Cornell Institute for Social and Economic Research (CISER), Cornell University, Ithaca NY<ExtLink URI="http://ciser
20 20      </prodPlac>
21 21      <software>CED2AR, Version 1.0</software>
22 22      + <software>The Comprehensive Extensible Data Documentation and Access Repository 2.5</software>
23 23      <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
24 24      <grantNo agency="National Science Foundation">1131848</grantNo>
25 25      </prodStmt>

```



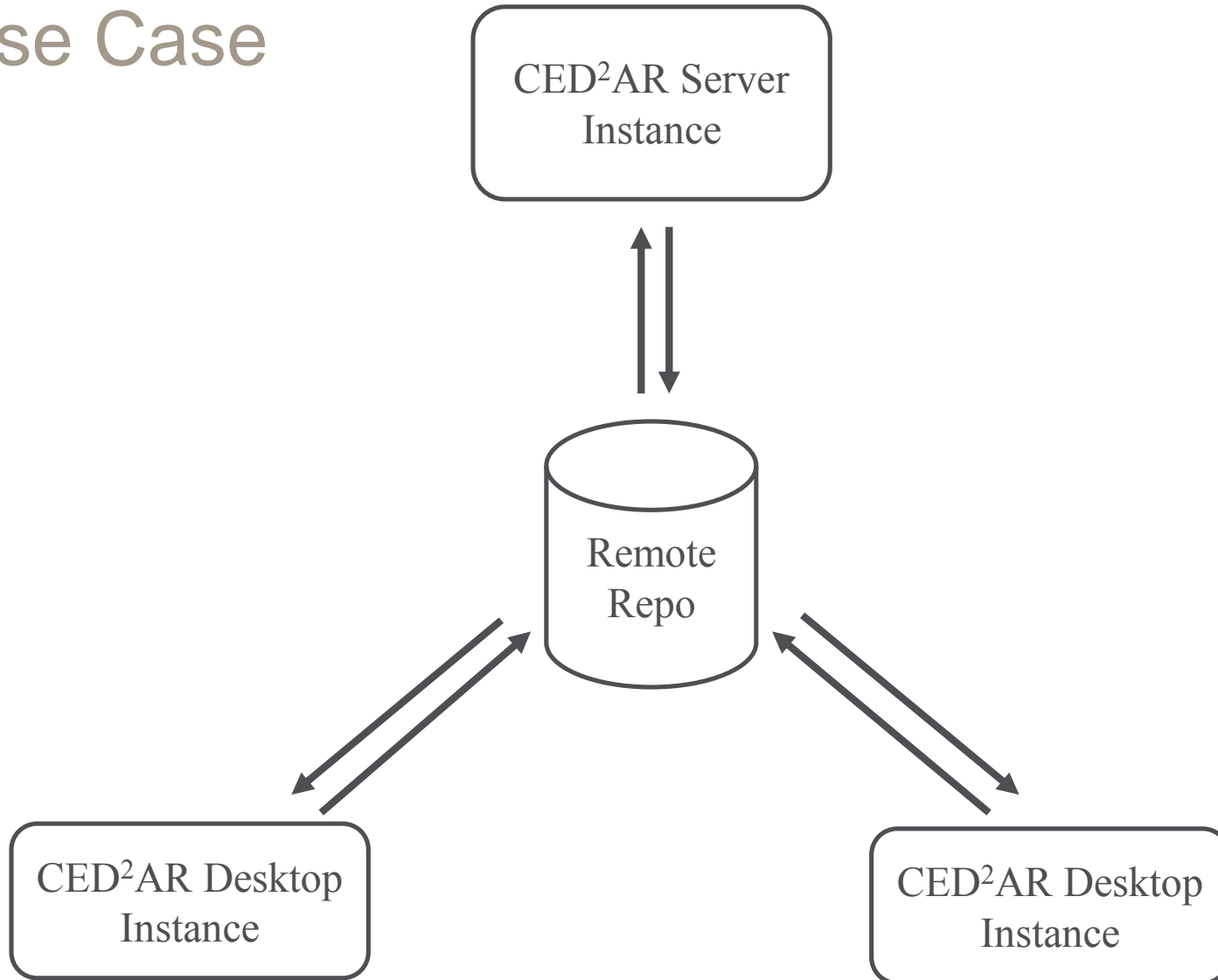
```

26 26      <distDate date="2014">2014</distDate>
27 27      </distStmt>
28 28      <verStmt>
29 29      - <version date="2014-10-07 09:10:40 (auto-generated)">2014-06-18</version>
30 30      + <version date="2014-11-13 10:38:45 (auto-generated)">2014-06-18</version>
31 31      </verStmt>
32 32      <biblCit>Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook
33 33      </citation>

```



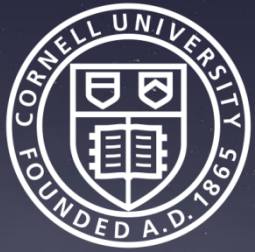
Use Case





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