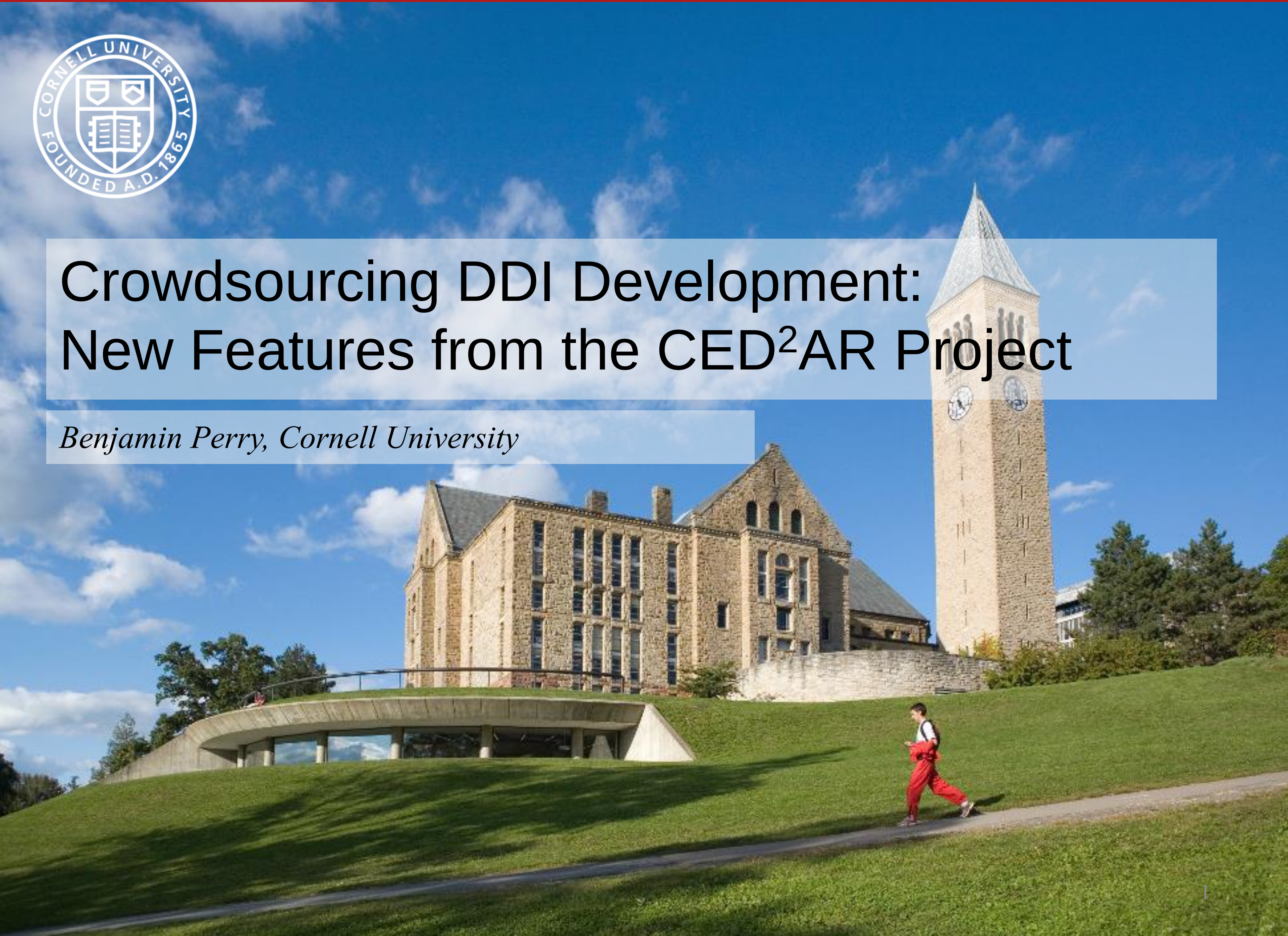
Crowdsourcing DDI Development:
New Features from the CED^{2}AR Project

Benjamin Perry, Cornell University



What is CED^{2}AR?

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

Initial Focus

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

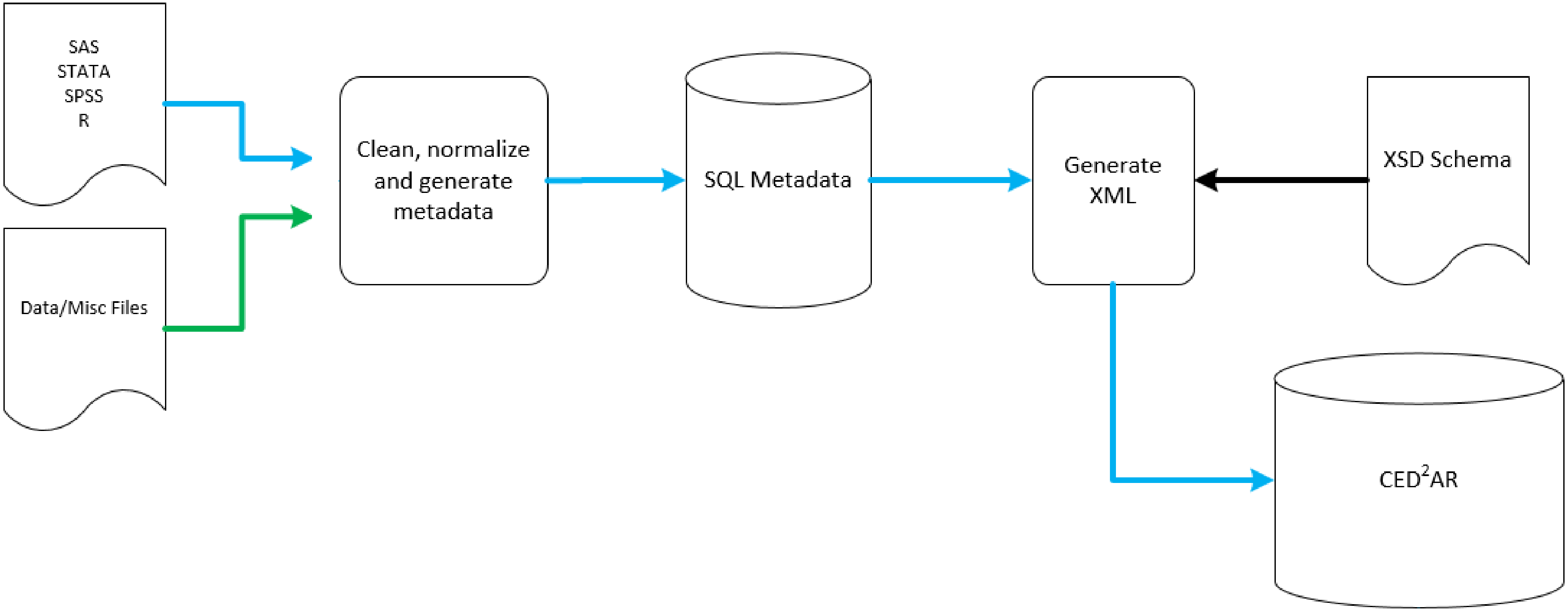
Now

- 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

Starting point here

- 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

Ingest Workflow



User Workflow

1. User searches through CED^{2}AR 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

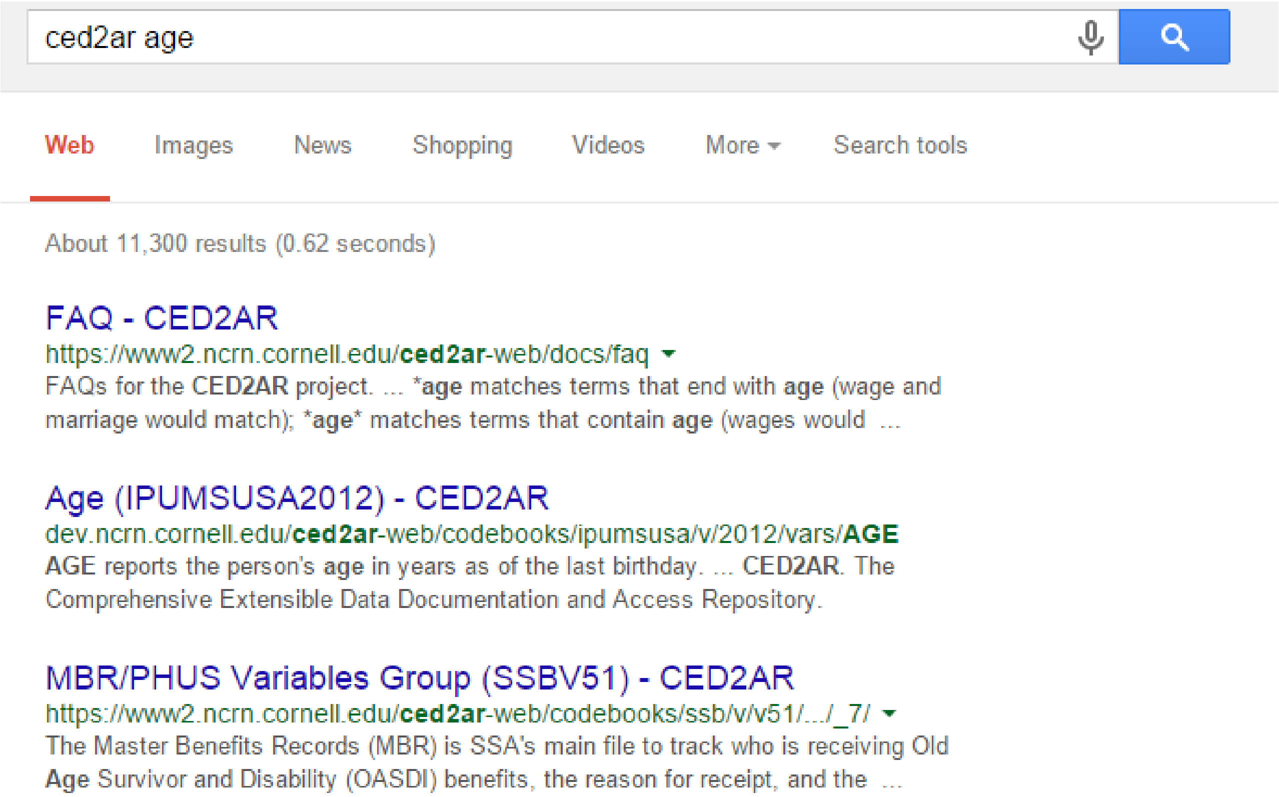
1. Search engine optimization enhancements to DDI
2. Exposing community contributions

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



About 11,300 results (0.62 seconds)

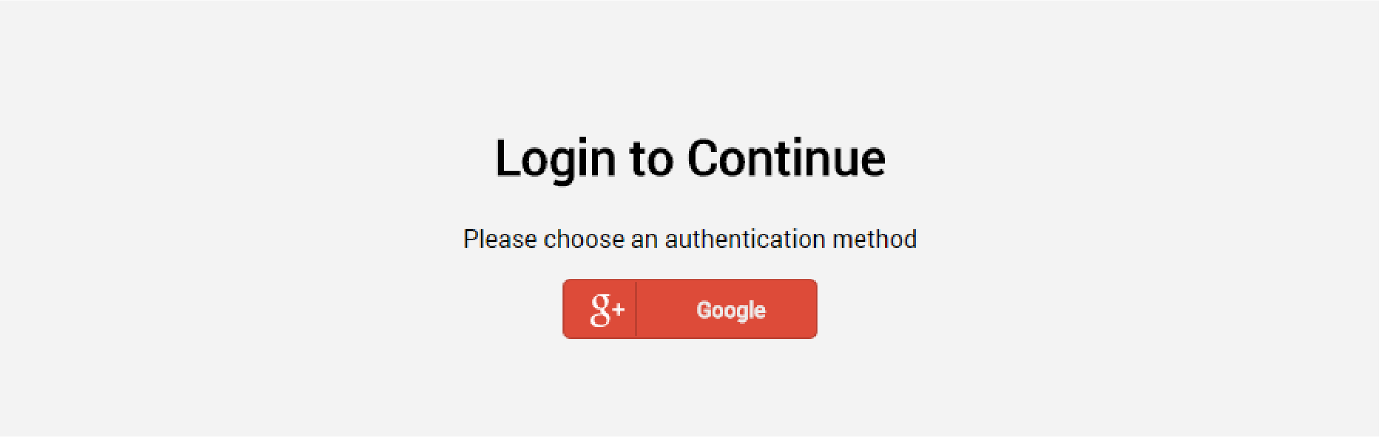
FAQ - CED2AR
https://www2.ncrn.cornell.edu/**ced2ar**-web/docs/faq
FAQs for the **CED2AR** project. ... ***age** matches terms that end with **age** (wage and marriage would match); ***age*** matches terms that contain **age** (wages would ...

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/ssb/v/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 – currently working on ORCID
- CED^{2}AR handles identity management

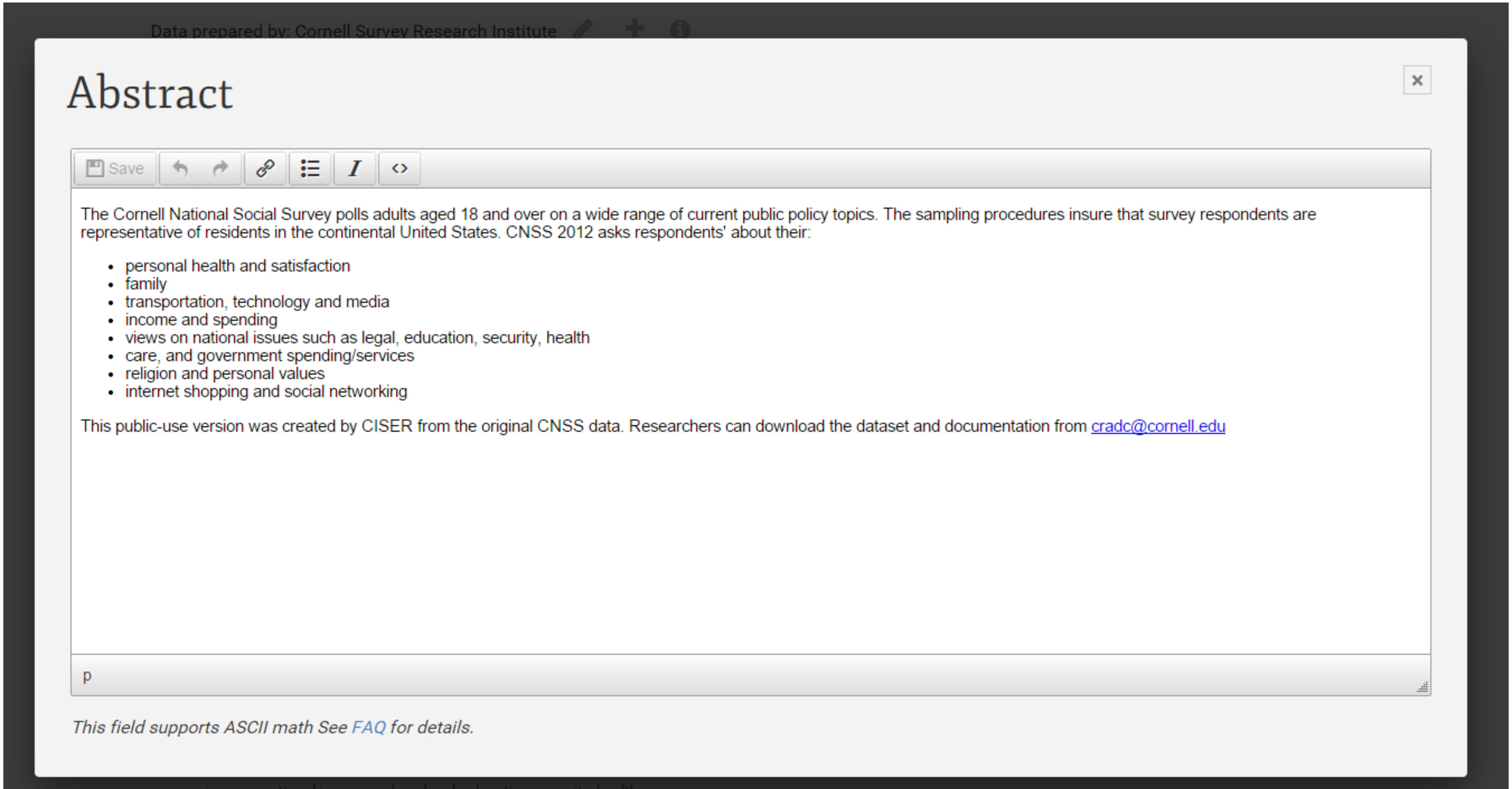


Login to Continue

Please choose an authentication method

Editing

- Automatic validation, and editor for rich content



Abstract

The Cornell National Social Survey polls adults aged 18 and over on a wide range of current public policy topics. The sampling procedures insure that survey respondents are representative of residents in the continental United States. CNSS 2012 asks respondents' about their:

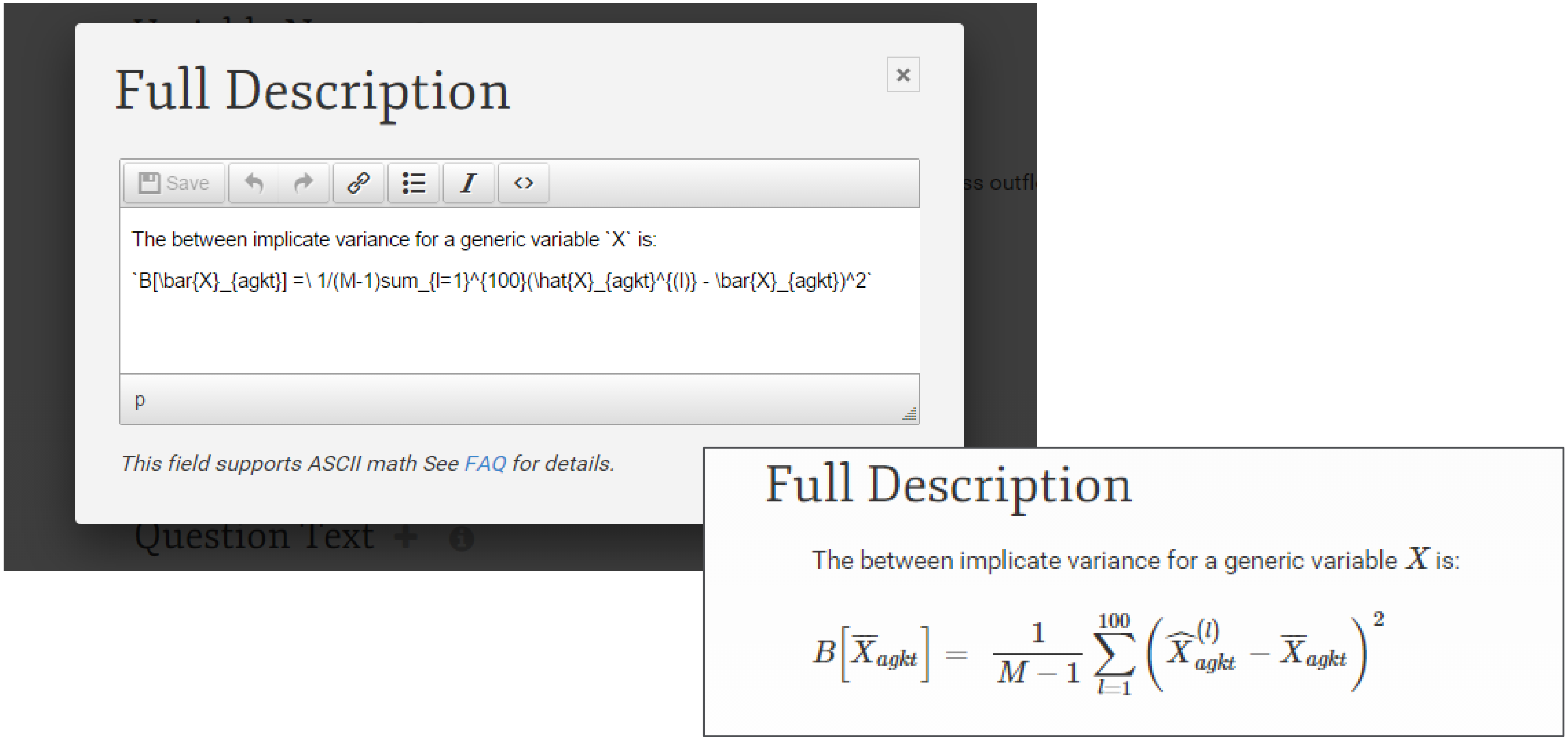
- personal health and satisfaction
- family
- transportation, technology and media
- income and spending
- views on national issues such as legal, education, security, health
- care, and government spending/services
- religion and personal values
- internet shopping and social networking

This public-use version was created by CISER from the original CNSS data. Researchers can download the dataset and documentation from cradc@cornell.edu

This field supports ASCII math See FAQ for details.

Editing

- Allows for ASCII Math



Full Description

```
The between implicate variance for a generic variable `X` is:
`B[\bar{X}_{agkt}] =\ 1/(M-1)sum_{l=1}^{100}(\hat{X}_{agkt}^{(l)} - \bar{X}_{agkt})^2`
```

p

This field supports ASCII math See FAQ for details.

Question Text

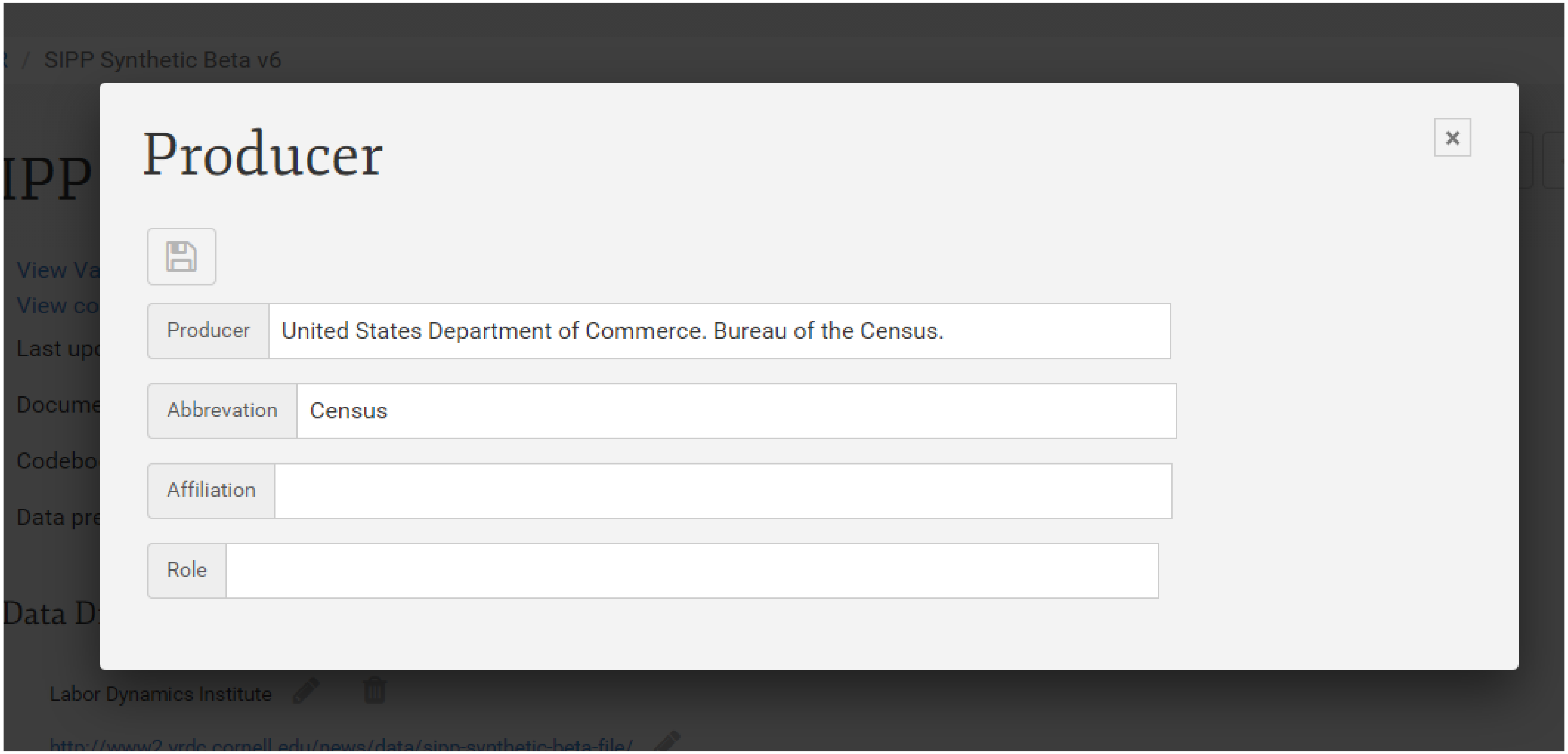
Full Description

The between implicate variance for a generic variable X is:

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

Editing

- Growing support for additional DDI fields, exposed or not

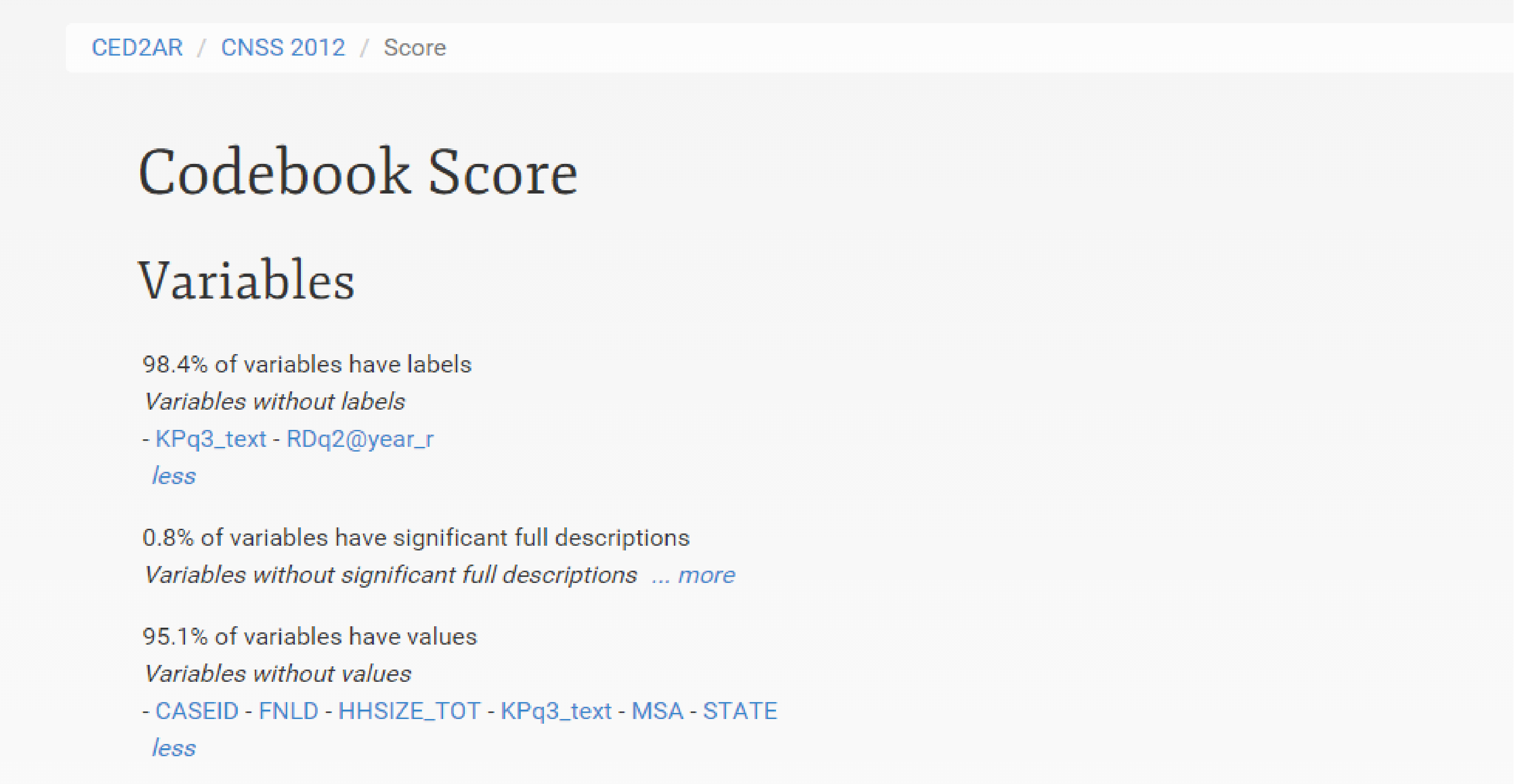


Producer

Producer	United States Department of Commerce. Bureau of the Census.
Abbrevation	Census
Affiliation	
Role	

Metadata Scoring

- Exposing sparse documentation



CED2AR / CNSS 2012 / Score

Codebook Score

Variables

98.4% of variables have labels

Variables without labels

- KPq3_text - RDq2@year_r

less

0.8% of variables have significant full descriptions

Variables without significant full descriptions ... more

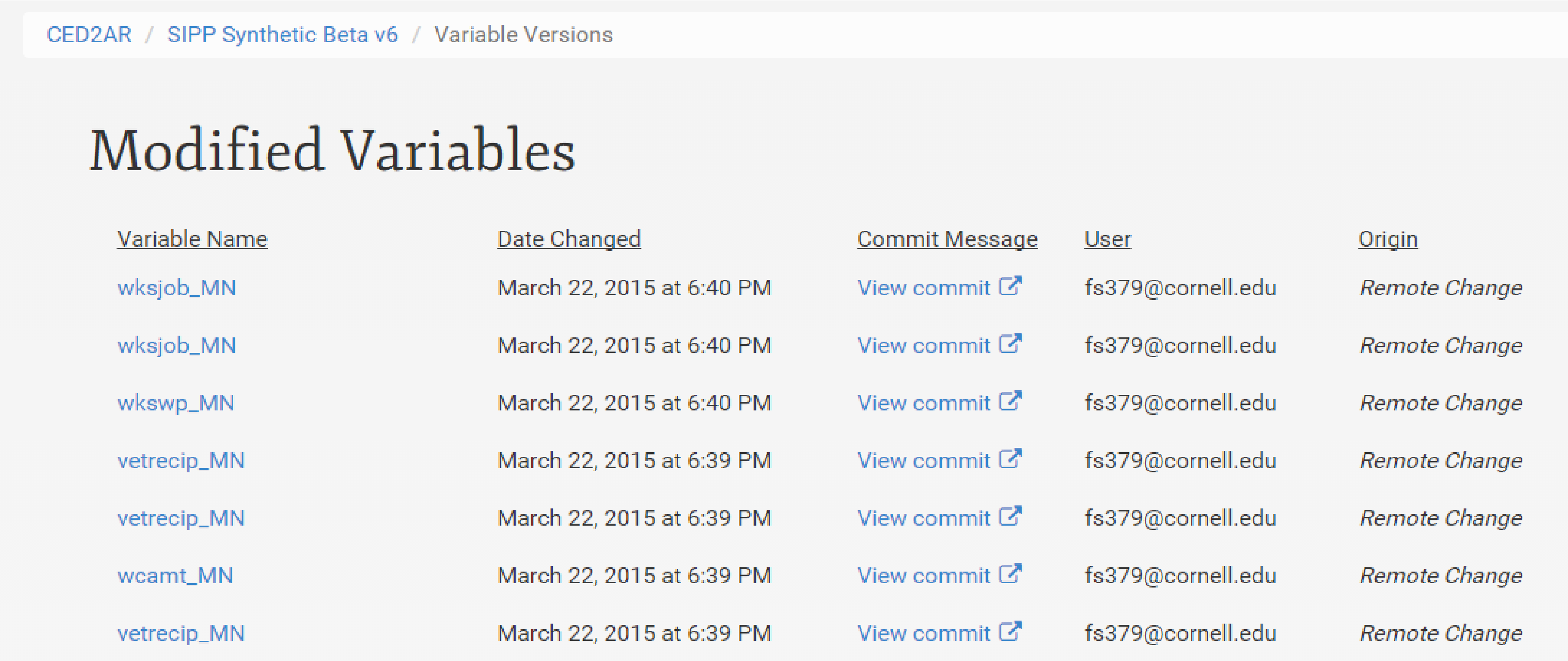
95.1% of variables have values

Variables without values

- CASEID - FNLD - HHSIZE_TOT - KPq3_text - MSA - STATE

less

User Contributions



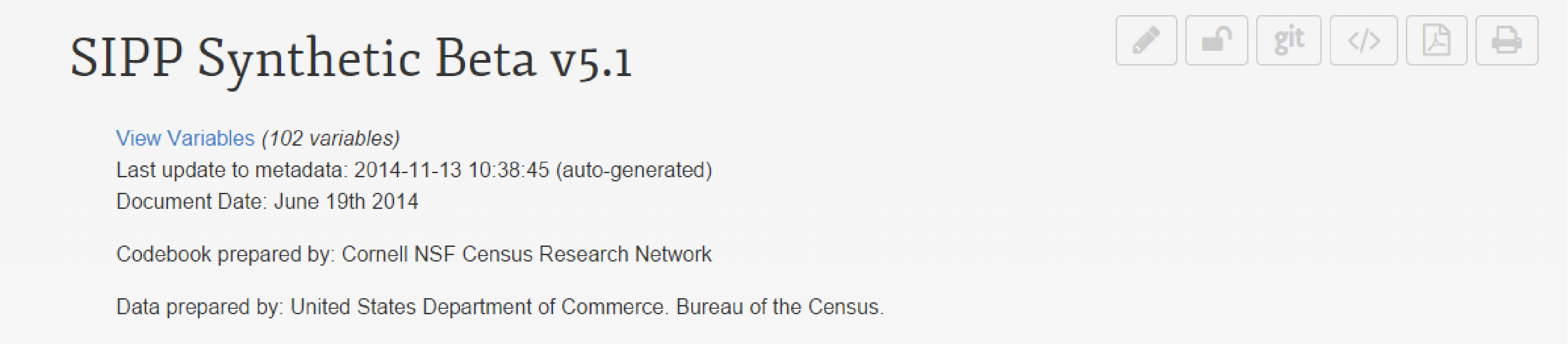
CED2AR / SIPP Synthetic Beta v6 / Variable Versions

Modified Variables

Variable Name	Date Changed	Commit Message	User	Origin
wksjob_MN	March 22, 2015 at 6:40 PM	View commit	fs379@cornell.edu	*Remote Change*
wksjob_MN	March 22, 2015 at 6:40 PM	View commit	fs379@cornell.edu	*Remote Change*
wkswp_MN	March 22, 2015 at 6:40 PM	View commit	fs379@cornell.edu	*Remote Change*
vetrecip_MN	March 22, 2015 at 6:39 PM	View commit	fs379@cornell.edu	*Remote Change*
vetrecip_MN	March 22, 2015 at 6:39 PM	View commit	fs379@cornell.edu	*Remote Change*
wcamt_MN	March 22, 2015 at 6:39 PM	View commit	fs379@cornell.edu	*Remote Change*
vetrecip_MN	March 22, 2015 at 6:39 PM	View commit	fs379@cornell.edu	*Remote Change*

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



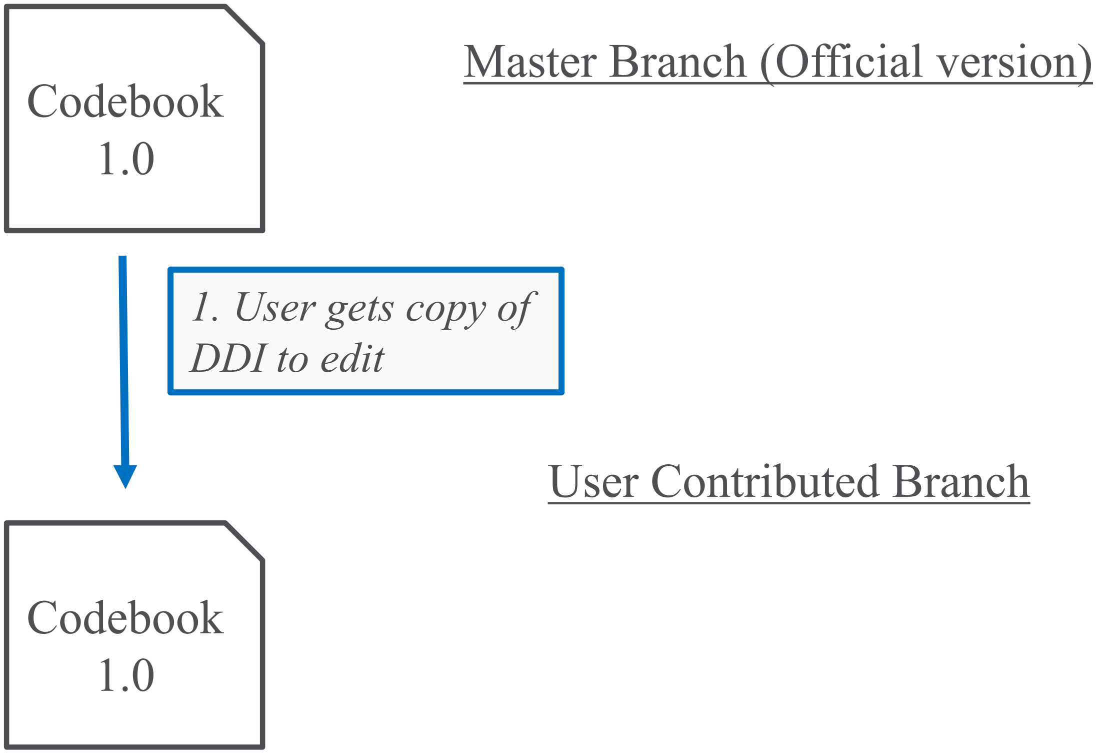
SIPP Synthetic Beta v5.1

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.

Architecture



Codebook 1.0

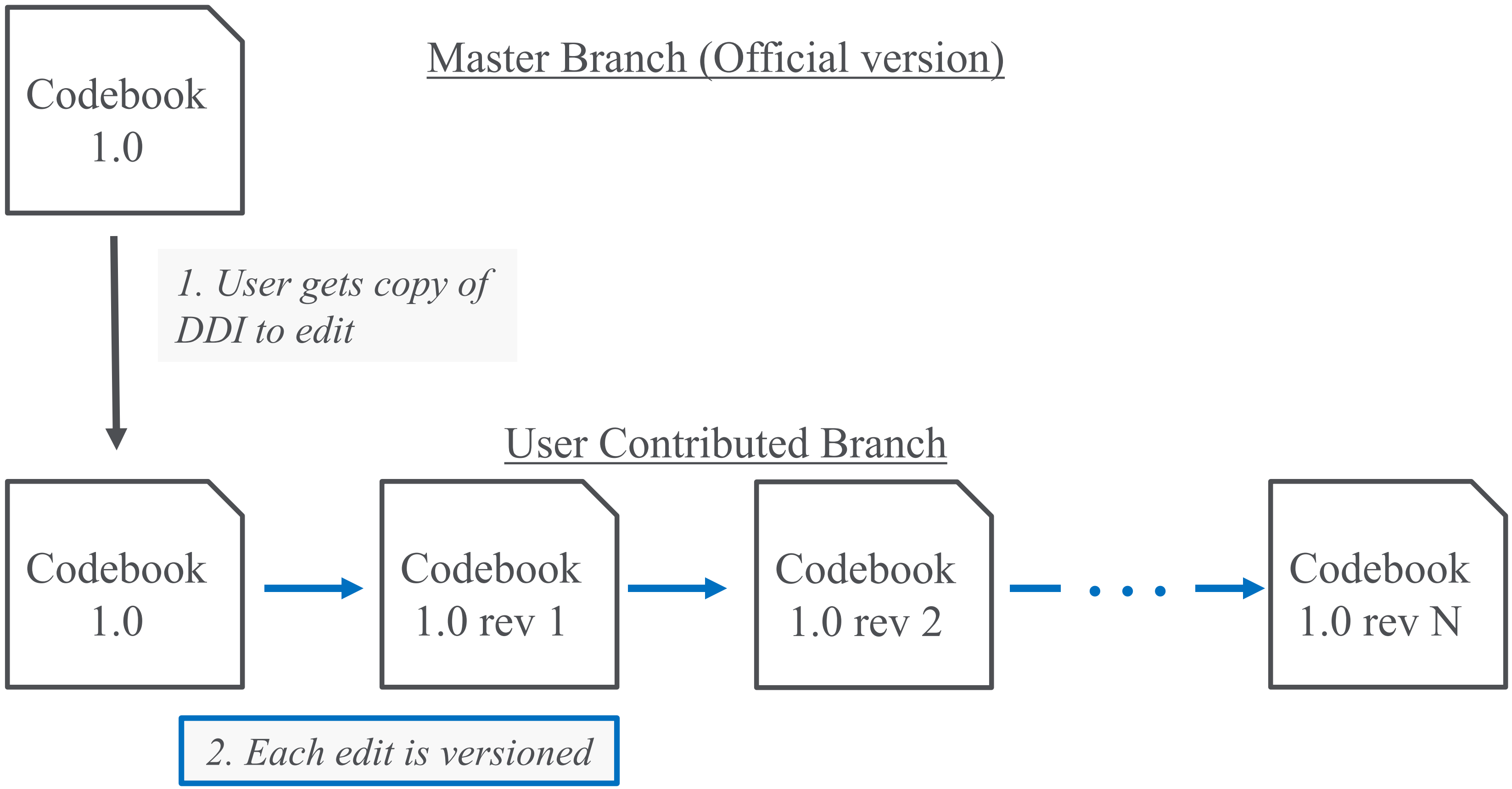
Master Branch (Official version)

1. User gets copy of DDI to edit

User Contributed Branch

Codebook 1.0

Architecture



Codebook 1.0

Master Branch (Official version)

1. User gets copy of DDI to edit

User Contributed Branch

Codebook 1.0

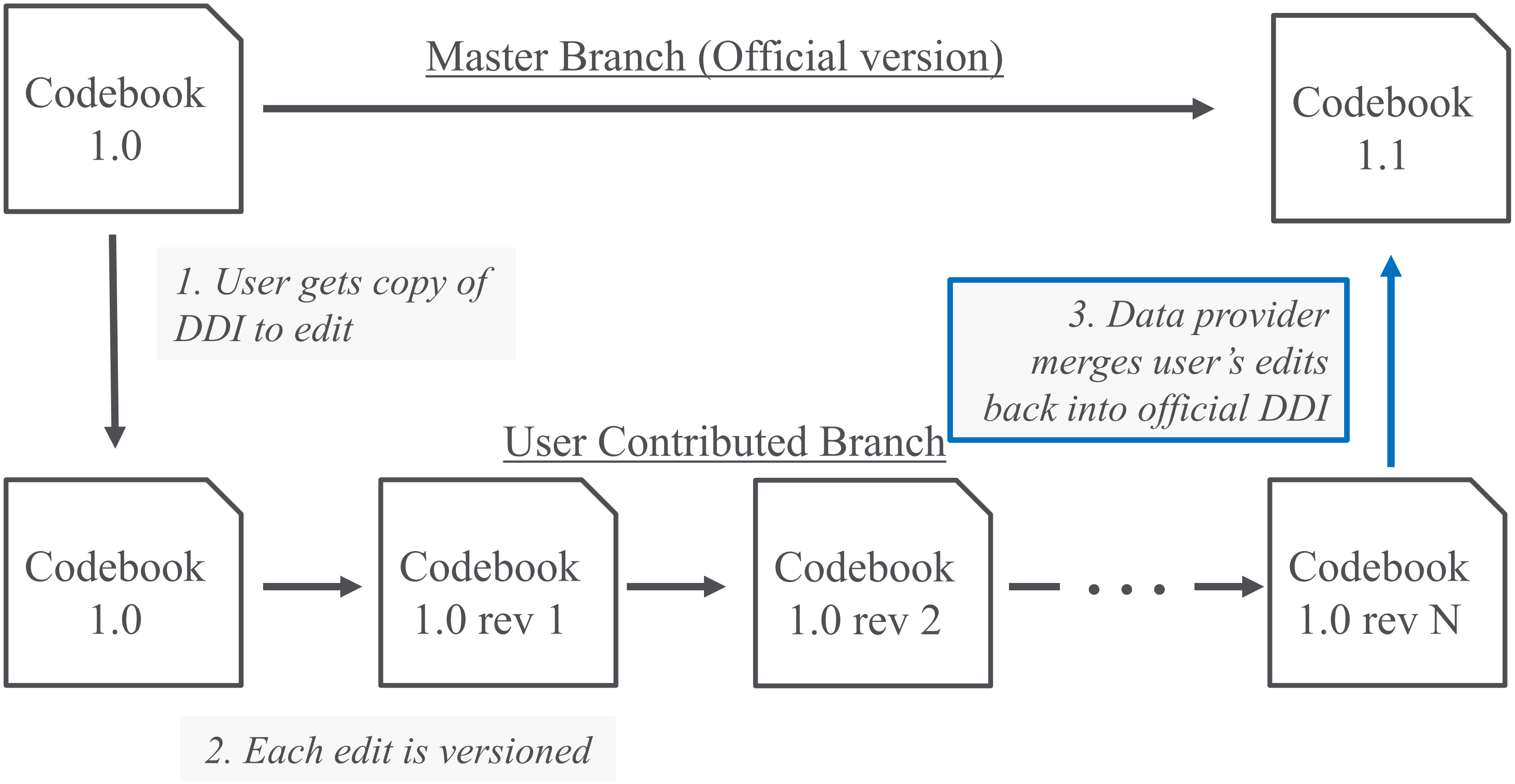
Codebook 1.0 rev 1

Codebook 1.0 rev 2

Codebook 1.0 rev N

2. Each edit is versioned

Architecture



Master Branch (Official version)

Codebook 1.0

Codebook 1.1

1. User gets copy of DDI to edit

User Contributed Branch

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

Codebook 1.0

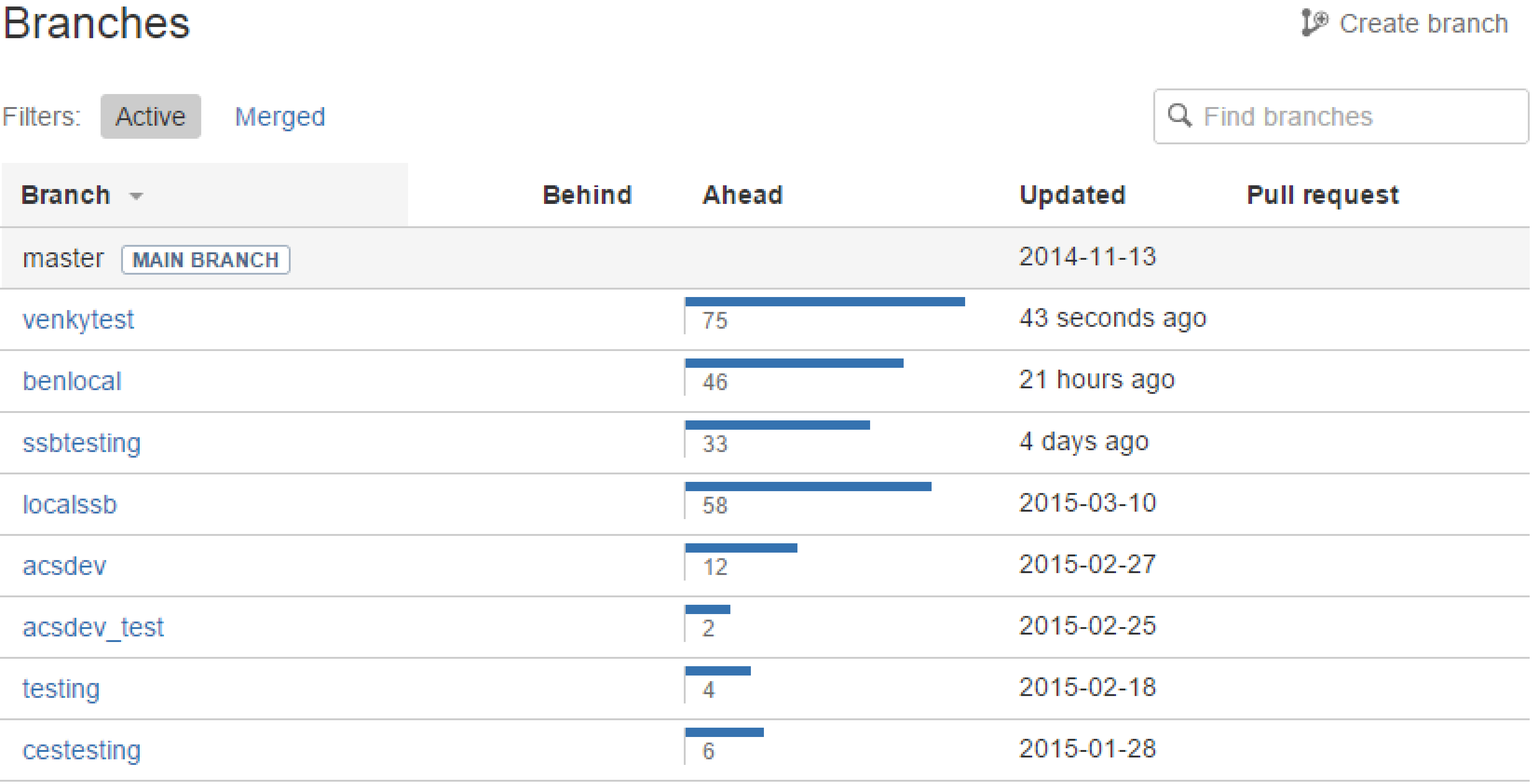
Codebook 1.0 rev 1

Codebook 1.0 rev 2

Codebook 1.0 rev N

2. Each edit is versioned

Architecture



Branches

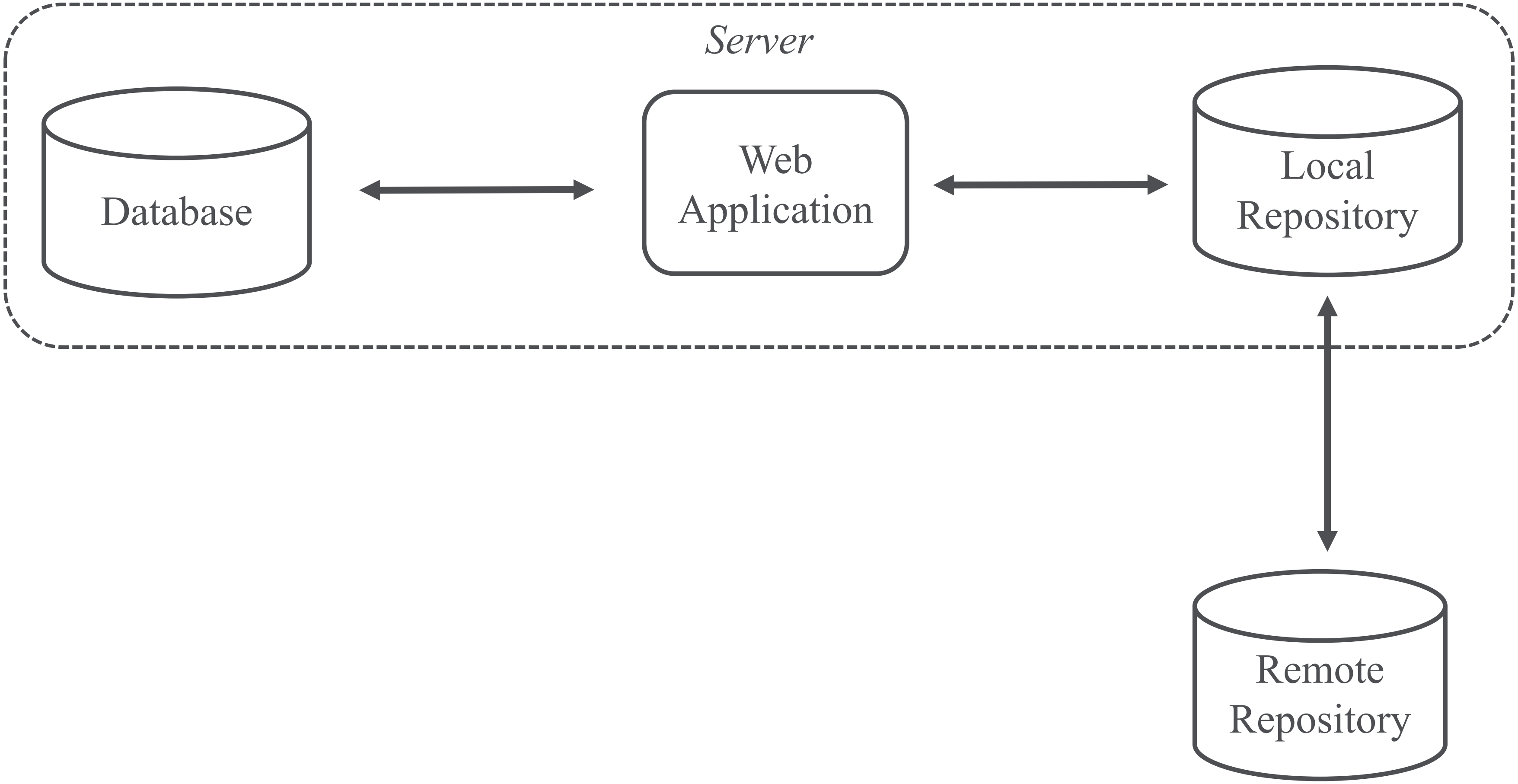
Create branch

Filters: Active Merged

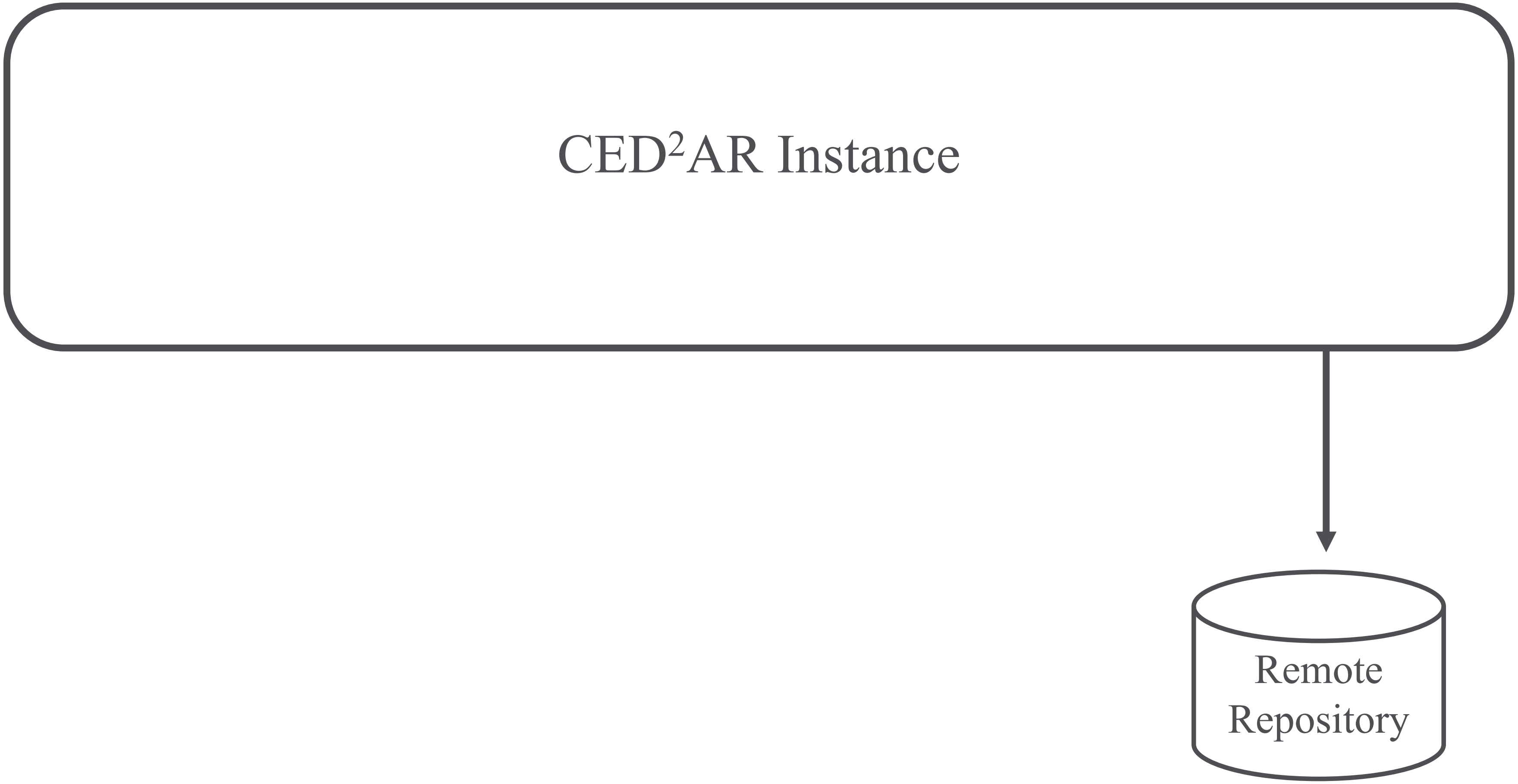
Find branches

Branch	Behind	Ahead	Updated	Pull request
master MAIN BRANCH			2014-11-13	
venkytest		75	43 seconds ago	
benlocal		46	21 hours ago	
ssbtesting		33	4 days ago	
localssb		58	2015-03-10	
acsdev		12	2015-02-27	
acsdev_test		2	2015-02-25	
testing		4	2015-02-18	
cestesting		6	2015-01-28	

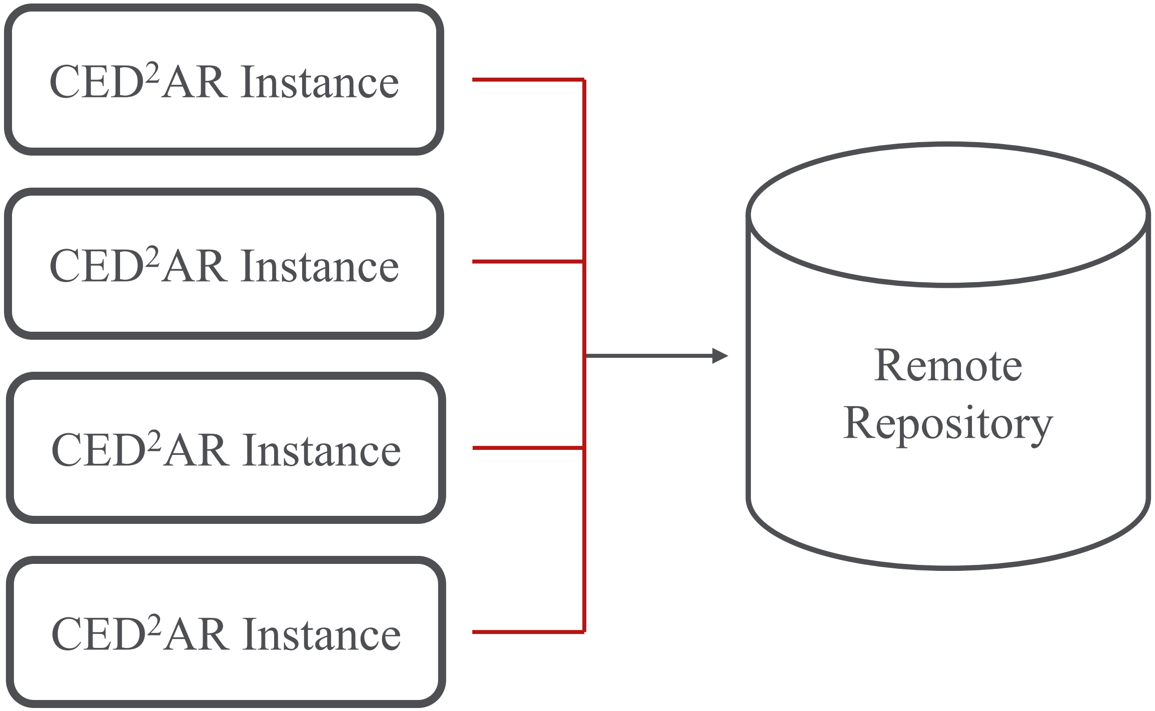
Architecture



Architecture



Architecture



CED^{2}AR Instance

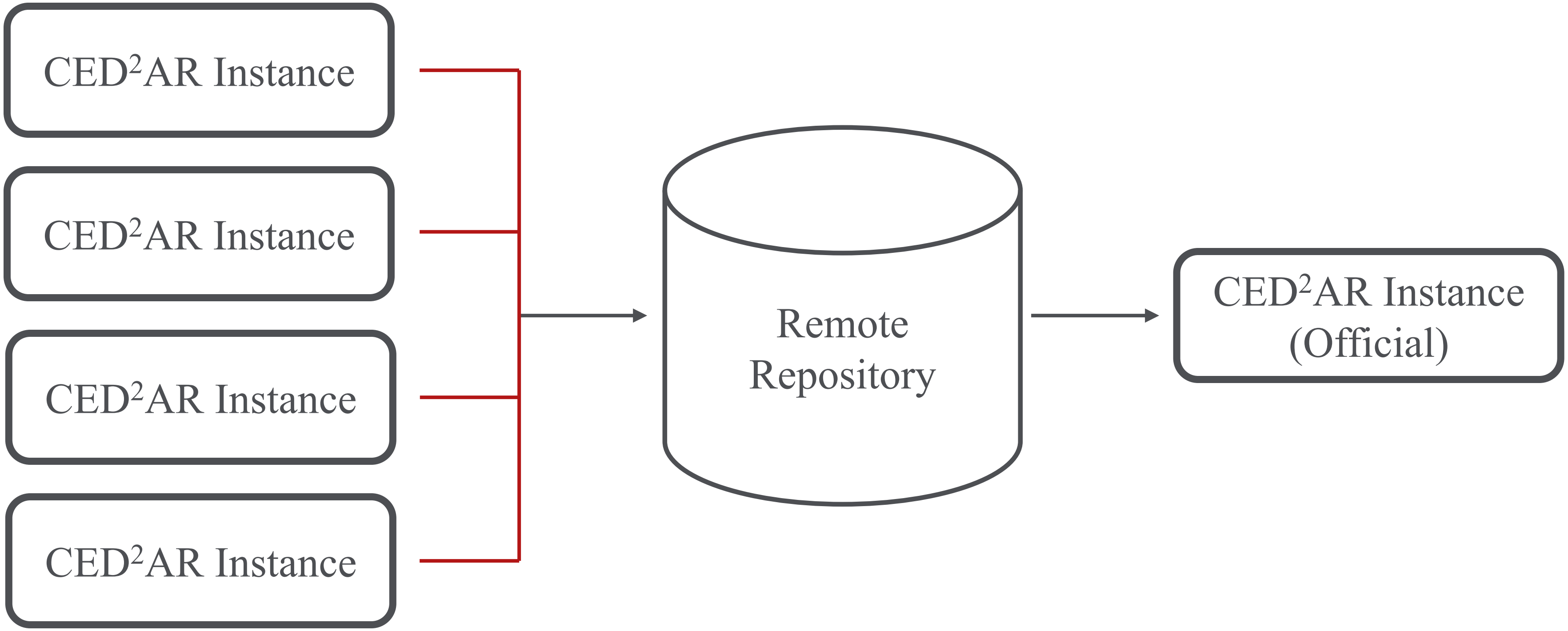
CED^{2}AR Instance

CED^{2}AR Instance

CED^{2}AR Instance

Remote Repository

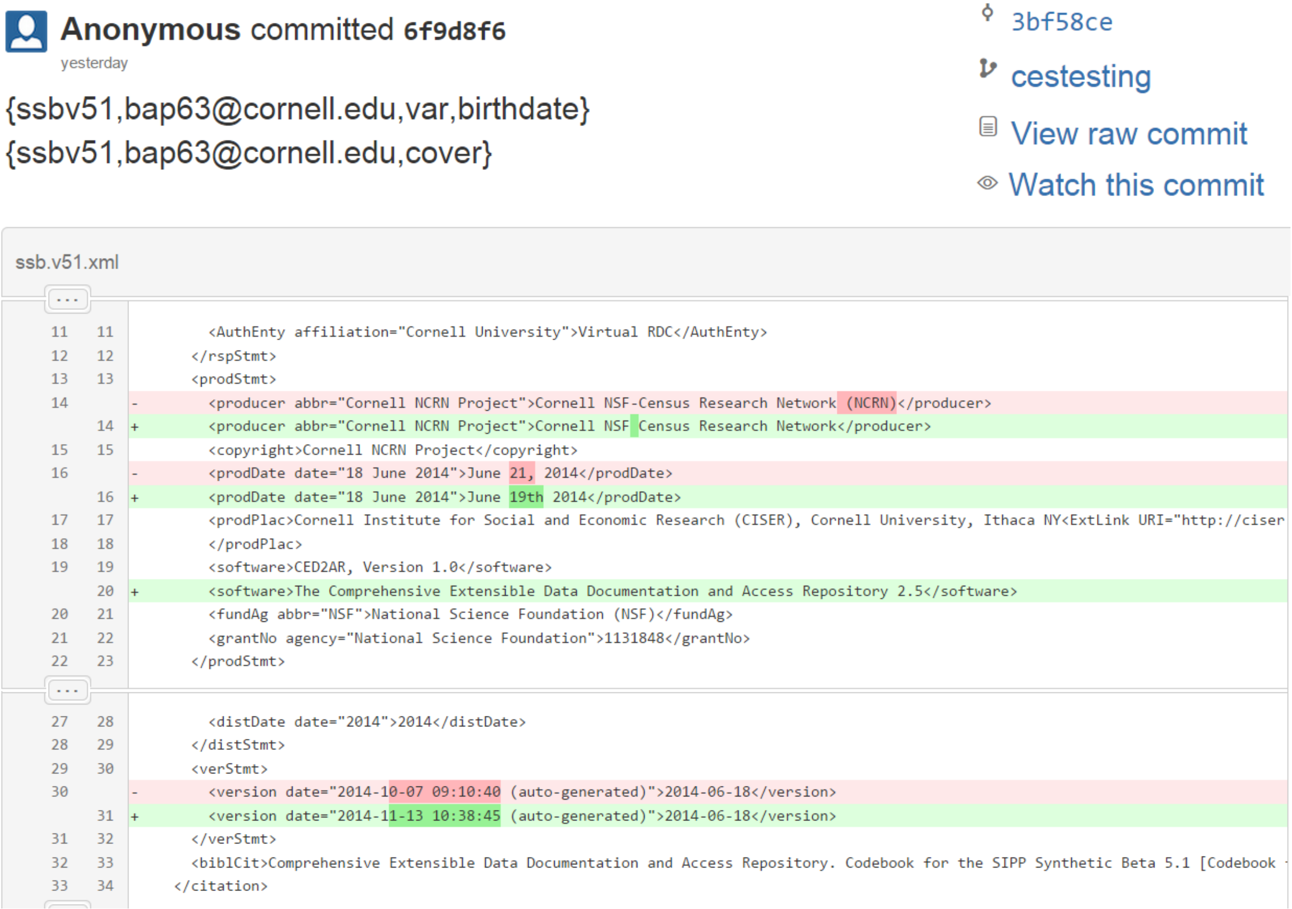
Architecture



Remote Location

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

Remote Location



Anonymous committed **6f9d8f6**

yesterday

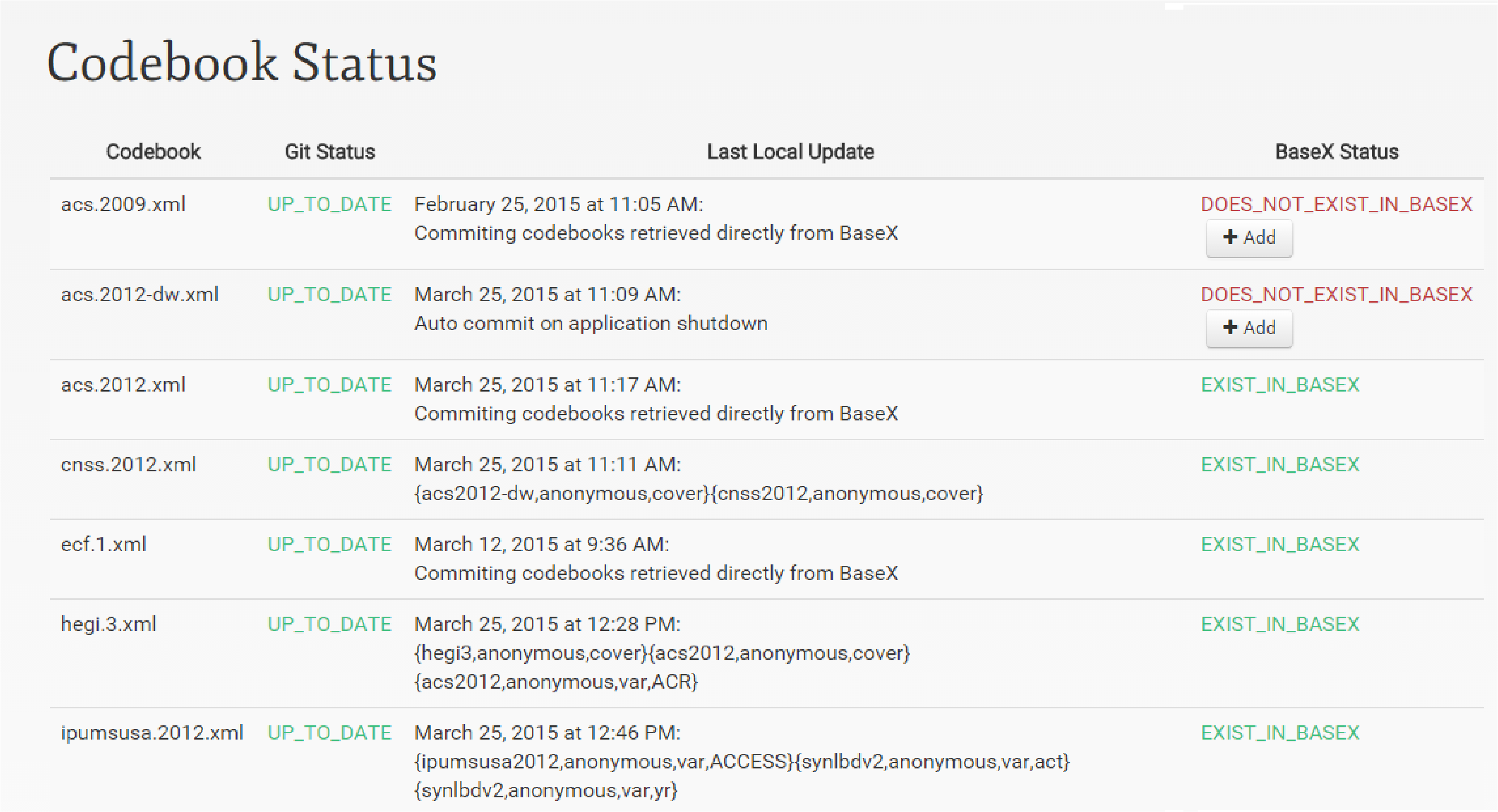
{ssbv51,bap63@cornell.edu,var,birthdate}
{ssbv51,bap63@cornell.edu,cover}

ssb.v51.xml

```
      <AuthEnty affiliation="Cornell University">Virtual RDC</AuthEnty>
    </rspStmt>
    <prodStmt>
-     <producer abbr="Cornell NCRN Project">Cornell NSF-Census Research Network (NCRN)</producer>
+     <producer abbr="Cornell NCRN Project">Cornell NSF Census Research Network</producer>
      <copyright>Cornell NCRN Project</copyright>
-     <prodDate date="18 June 2014">June 21, 2014</prodDate>
+     <prodDate date="18 June 2014">June 19th 2014</prodDate>
      <prodPlac>Cornell Institute for Social and Economic Research (CISER), Cornell University, Ithaca NY<ExtLink URI="http://ciser
      </prodPlac>
      <software>CED2AR, Version 1.0</software>
+     <software>The Comprehensive Extensible Data Documentation and Access Repository 2.5</software>
      <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
      <grantNo agency="National Science Foundation">1131848</grantNo>
    </prodStmt>
```

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

Tracking Changes



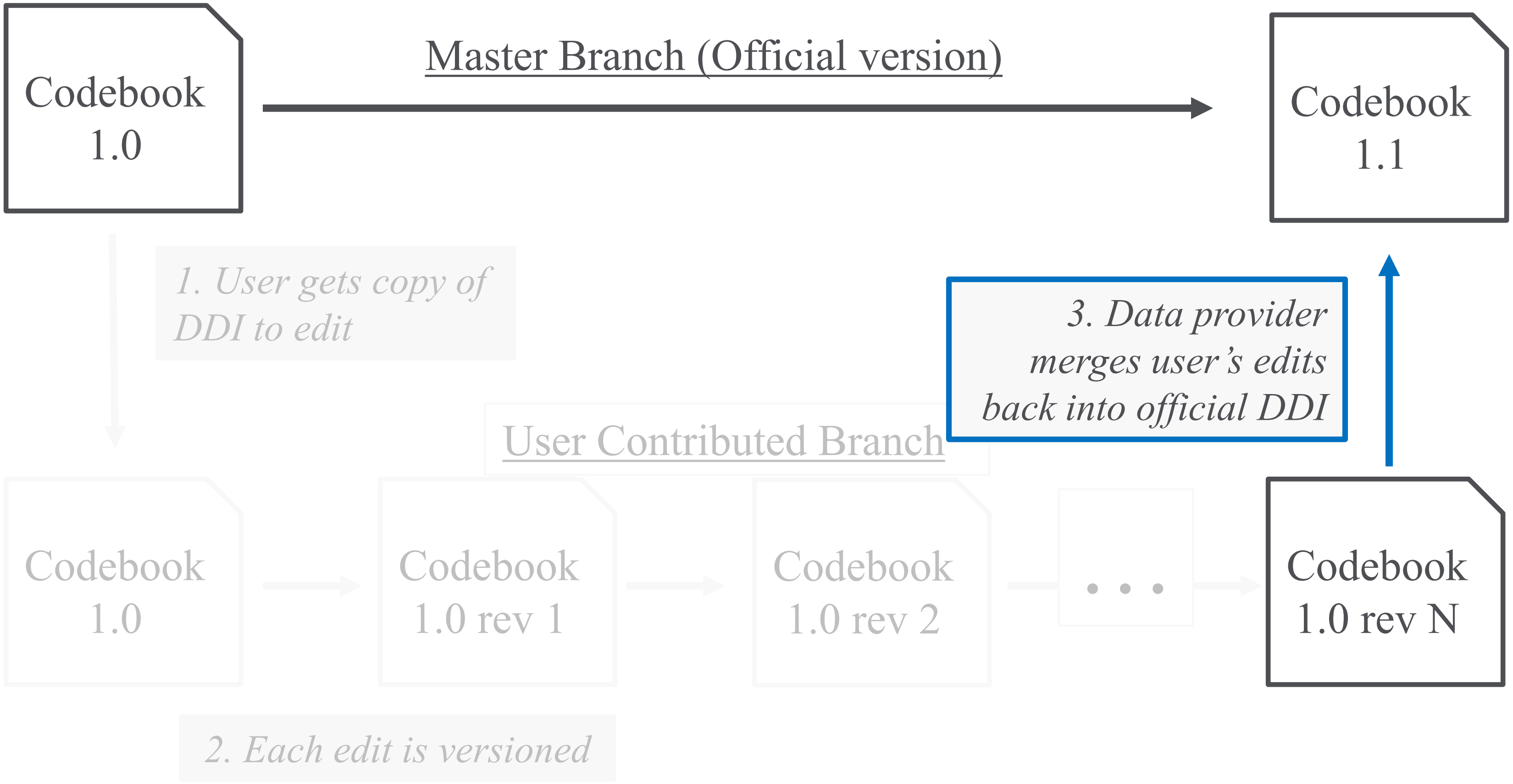
Codebook Status

Codebook	Git Status	Last Local Update	BaseX Status
acs.2009.xml	UP_TO_DATE	February 25, 2015 at 11:05 AM: Commiting codebooks retrieved directly from BaseX	DOES_NOT_EXIST_IN_BASEX + Add
acs.2012-dw.xml	UP_TO_DATE	March 25, 2015 at 11:09 AM: Auto commit on application shutdown	DOES_NOT_EXIST_IN_BASEX + Add
acs.2012.xml	UP_TO_DATE	March 25, 2015 at 11:17 AM: Commiting codebooks retrieved directly from BaseX	EXIST_IN_BASEX
cnss.2012.xml	UP_TO_DATE	March 25, 2015 at 11:11 AM: {acs2012-dw,anonymous,cover}{cnss2012,anonymous,cover}	EXIST_IN_BASEX
ecf.1.xml	UP_TO_DATE	March 12, 2015 at 9:36 AM: Commiting codebooks retrieved directly from BaseX	EXIST_IN_BASEX

hegi.3.xml	UP_TO_DATE	March 25, 2015 at 12:28 PM:
		{hegi3,anonymous,cover}{acs2012,anonymous,cover}
		{acs2012,anonymous,var,ACR}
			EXIST_IN_BASEX

ipumsusa.2012.xml	UP_TO_DATE	March 25, 2015 at 12:46 PM:
		{ipumsusa2012,anonymous,var,ACCESS}{synlbdv2,anonymous,var,act}
		{synlbdv2,anonymous,var,yr}
			EXIST_IN_BASEX

Continued Work: Improving merge control



Codebook 1.0

Master Branch (Official version)

Codebook 1.1

1. User gets copy of DDI to edit

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

User Contributed Branch

Codebook 1.0

Codebook 1.0 rev 1

Codebook 1.0 rev 2

. . .

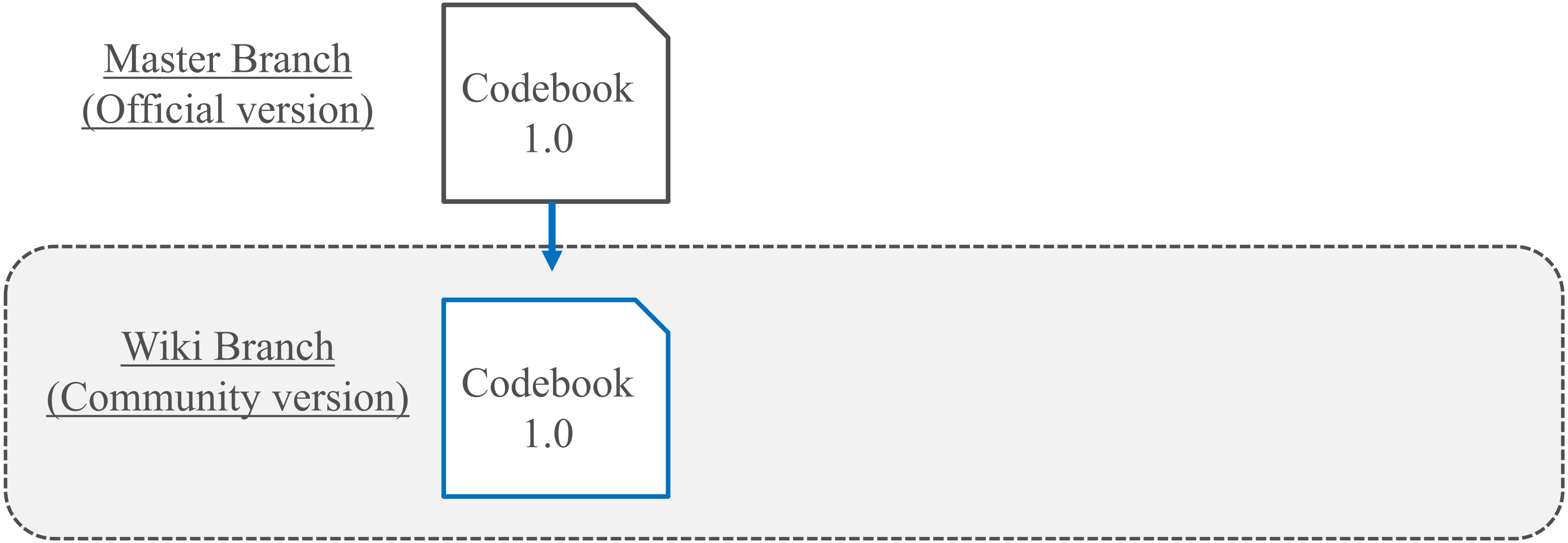
Codebook 1.0 rev N

2. Each edit is versioned

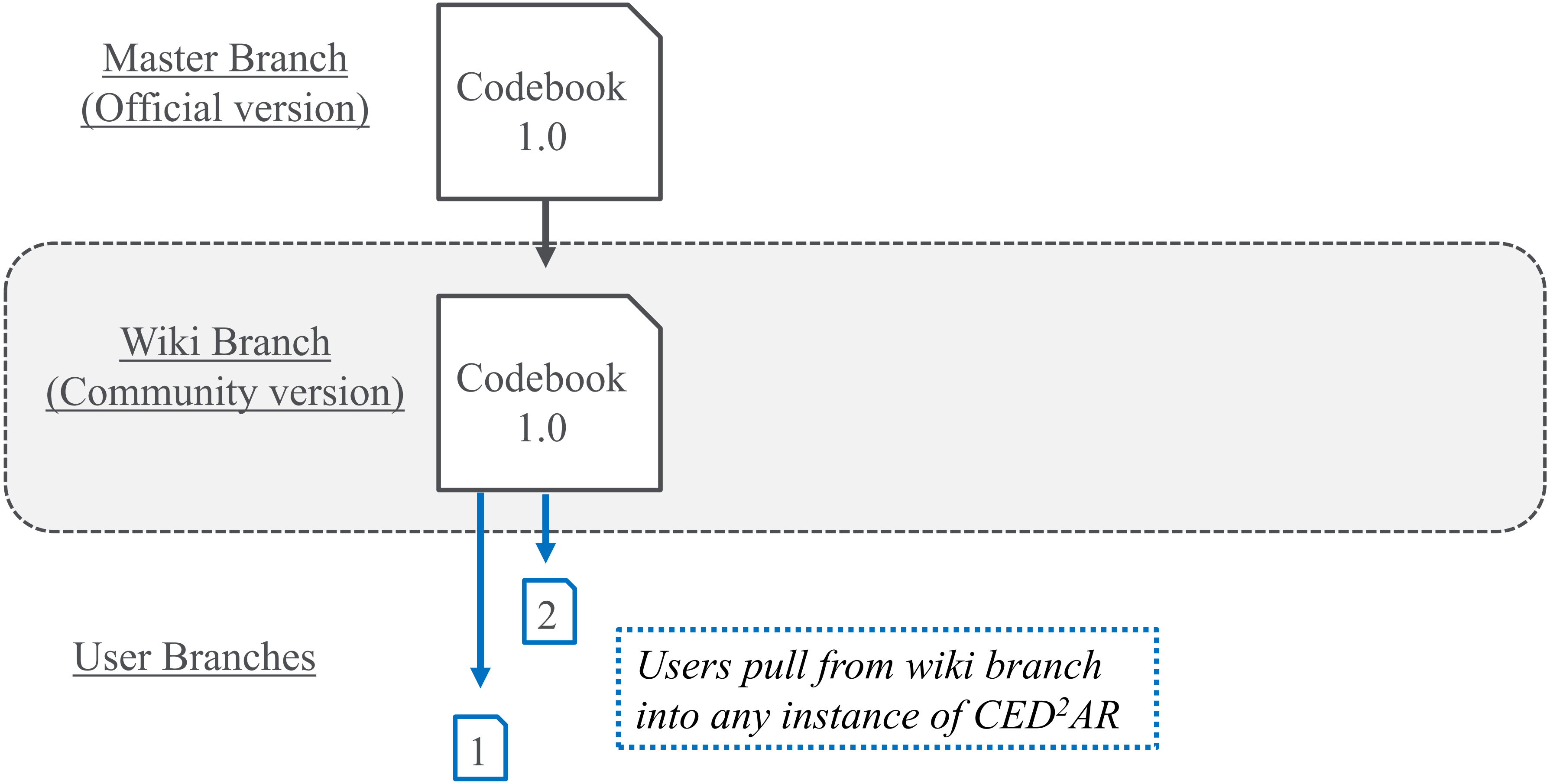
Continued Work: The uncontrolled merge

- 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)



Architecture (alternate)



Master Branch (Official version)

Codebook 1.0

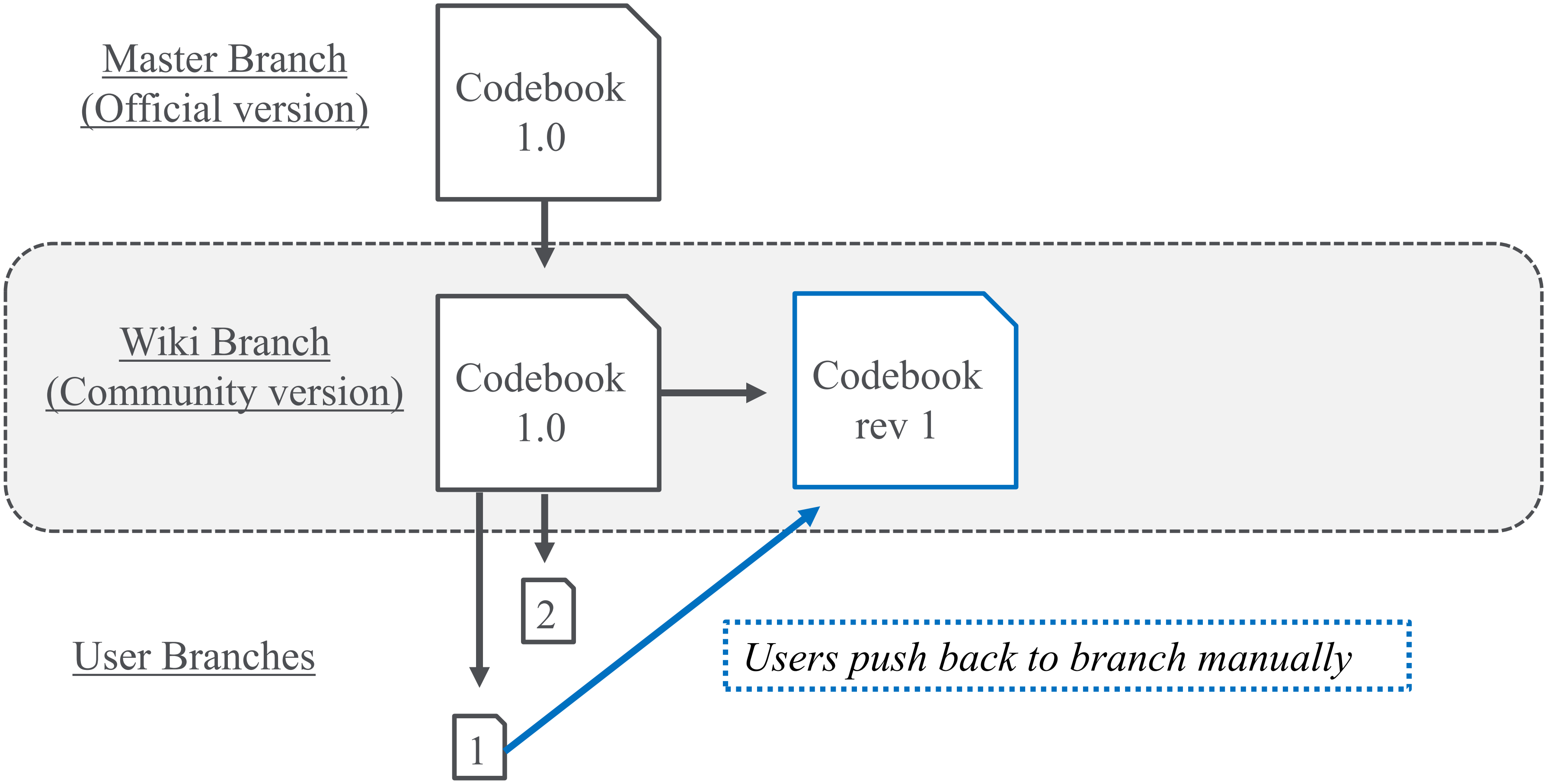
Wiki Branch (Community version)

Codebook 1.0

User Branches

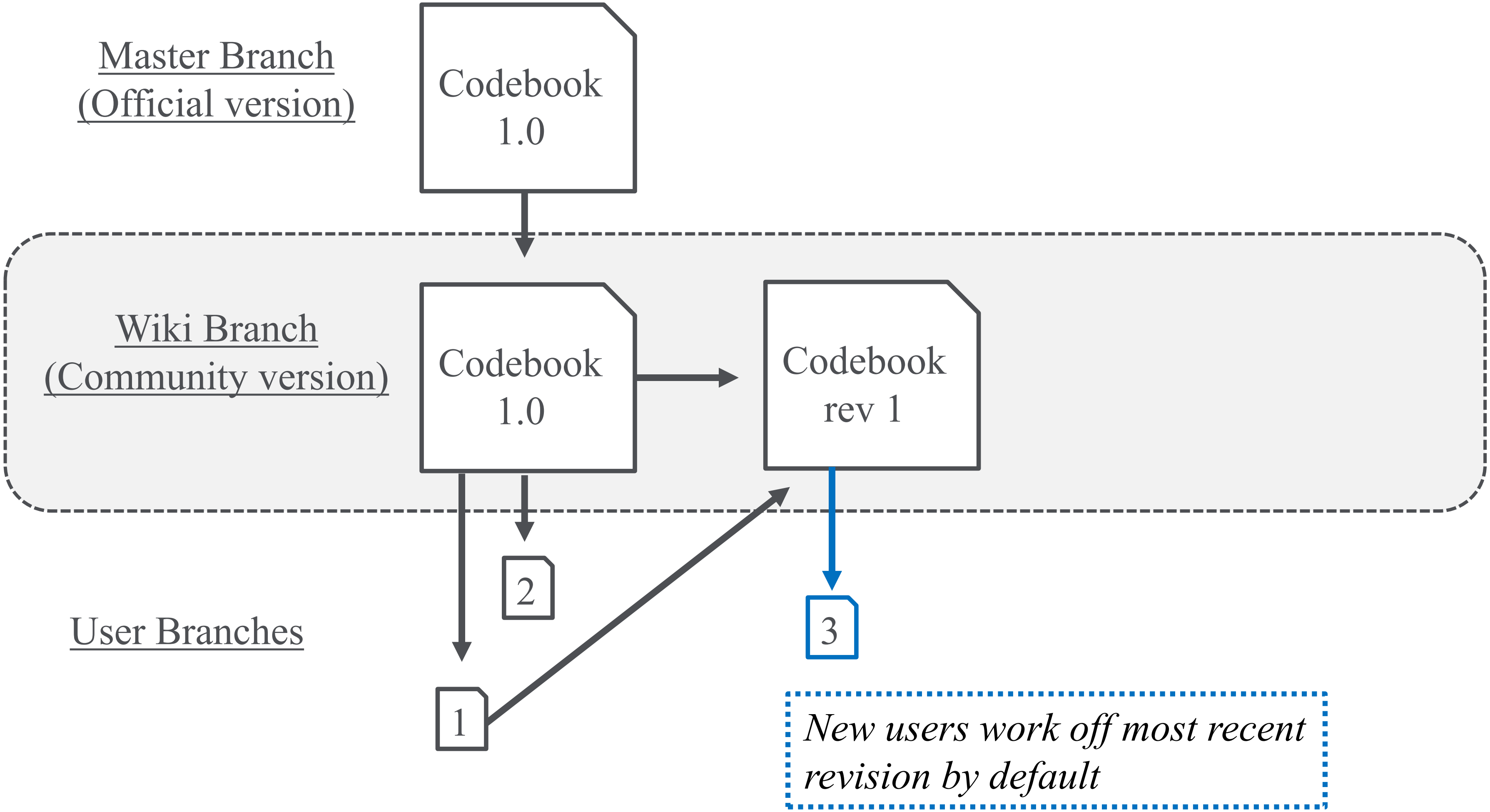
Users pull from wiki branch into any instance of CED^{2}AR

Architecture (alternate)



Users push back to branch manually

Architecture (alternate)



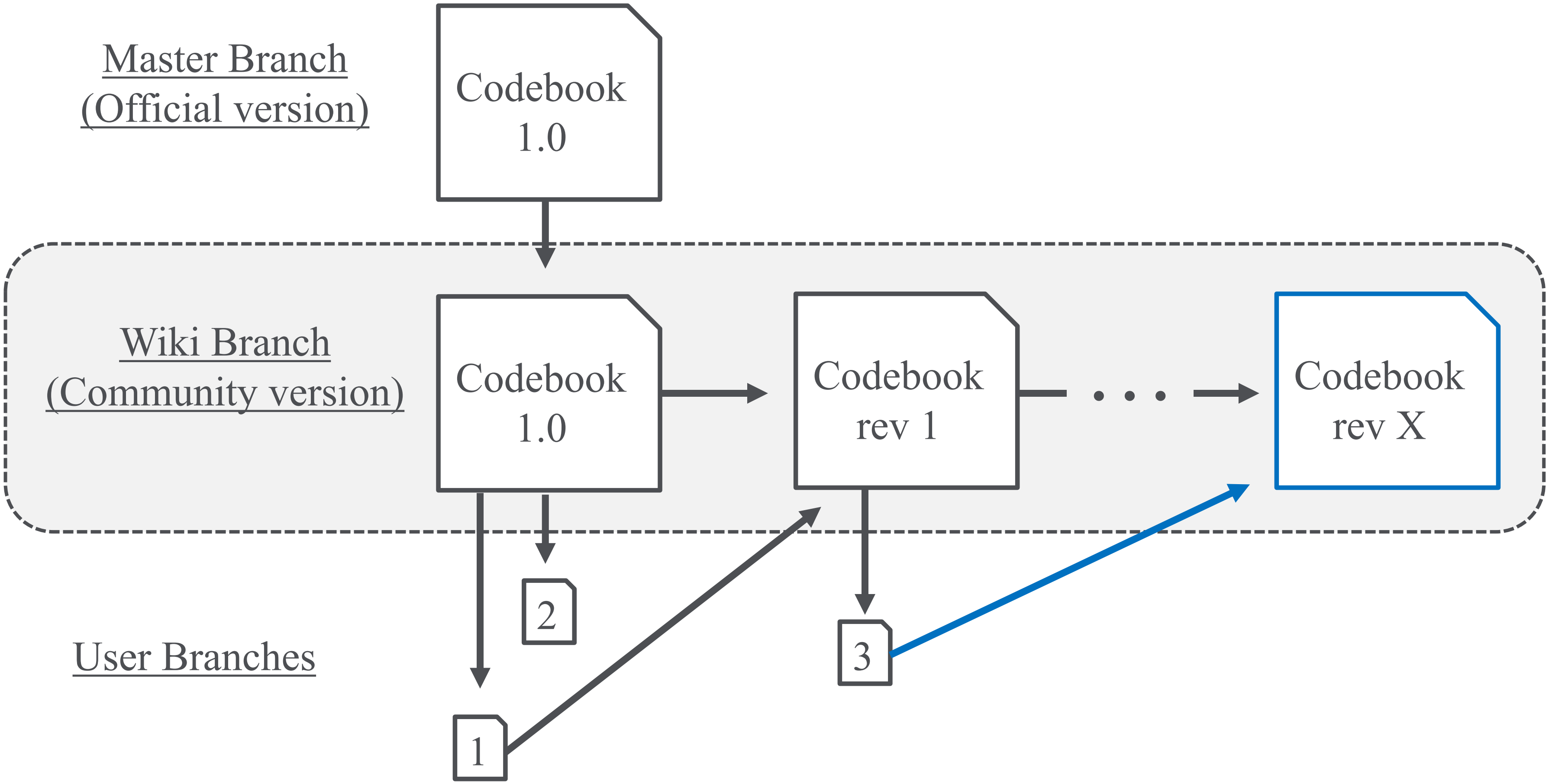
Master Branch
(Official version)

Wiki Branch
(Community version)

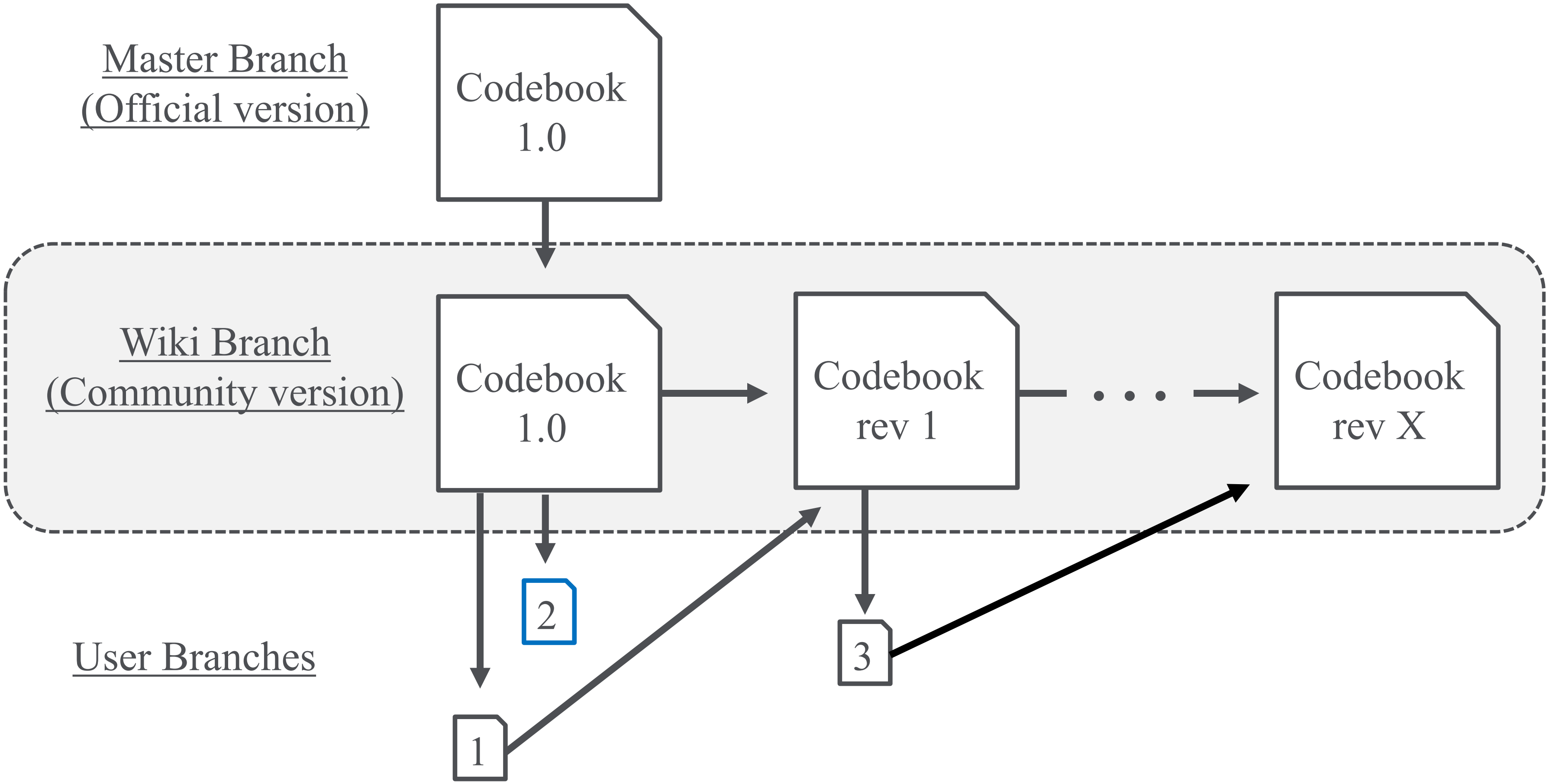
User Branches

New users work off most recent revision by default

Architecture (alternate)



Architecture (alternate)

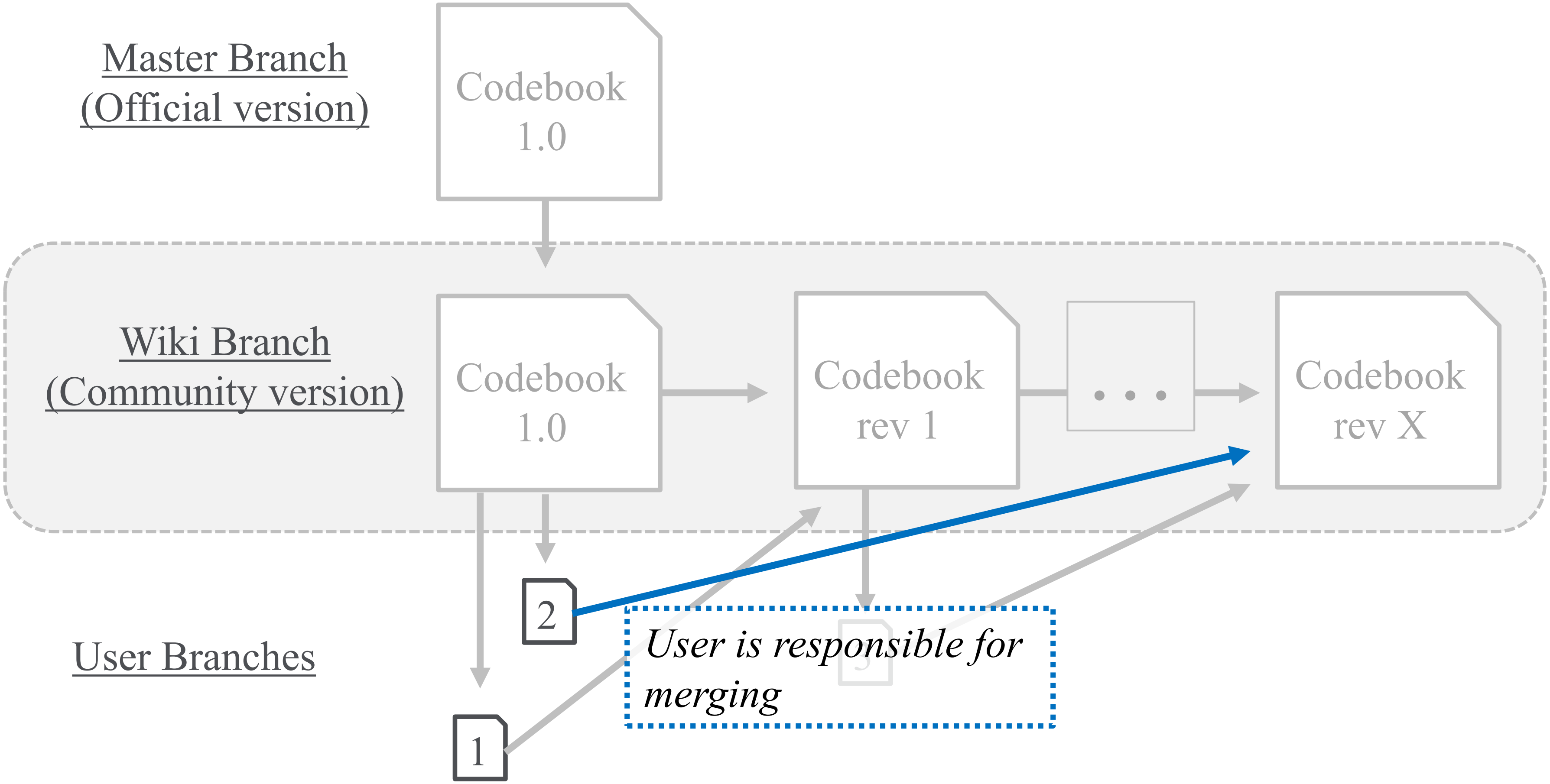


Master Branch (Official version)

Wiki Branch (Community version)

User Branches

Architecture (alternate)



Master Branch
(Official version)

Codebook
1.0

Wiki Branch
(Community version)

Codebook
1.0

Codebook
rev 1

. . .

Codebook
rev X

User Branches

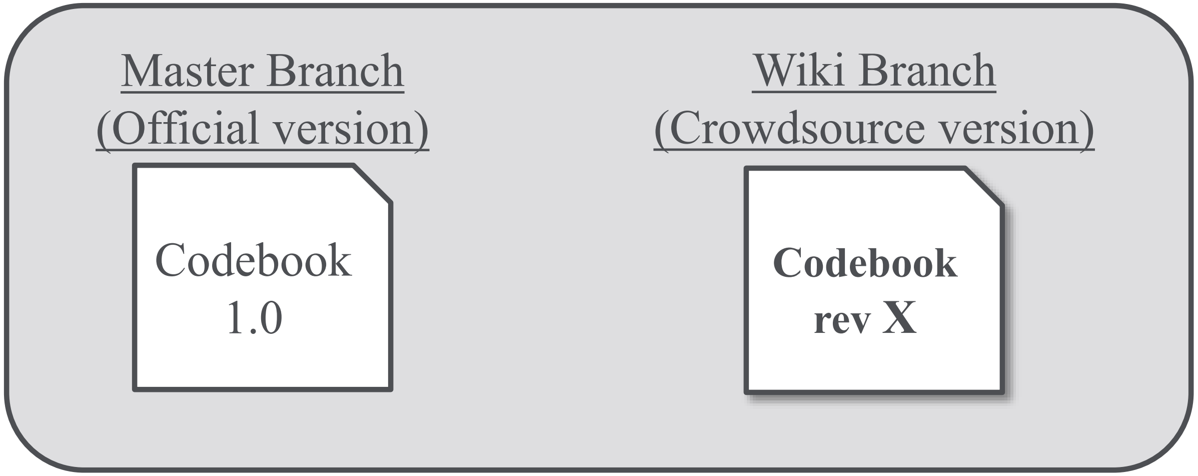
2

1

3

User is responsible for merging

Architecture (alternate)


Master Branch
(Official version)

Codebook
1.0

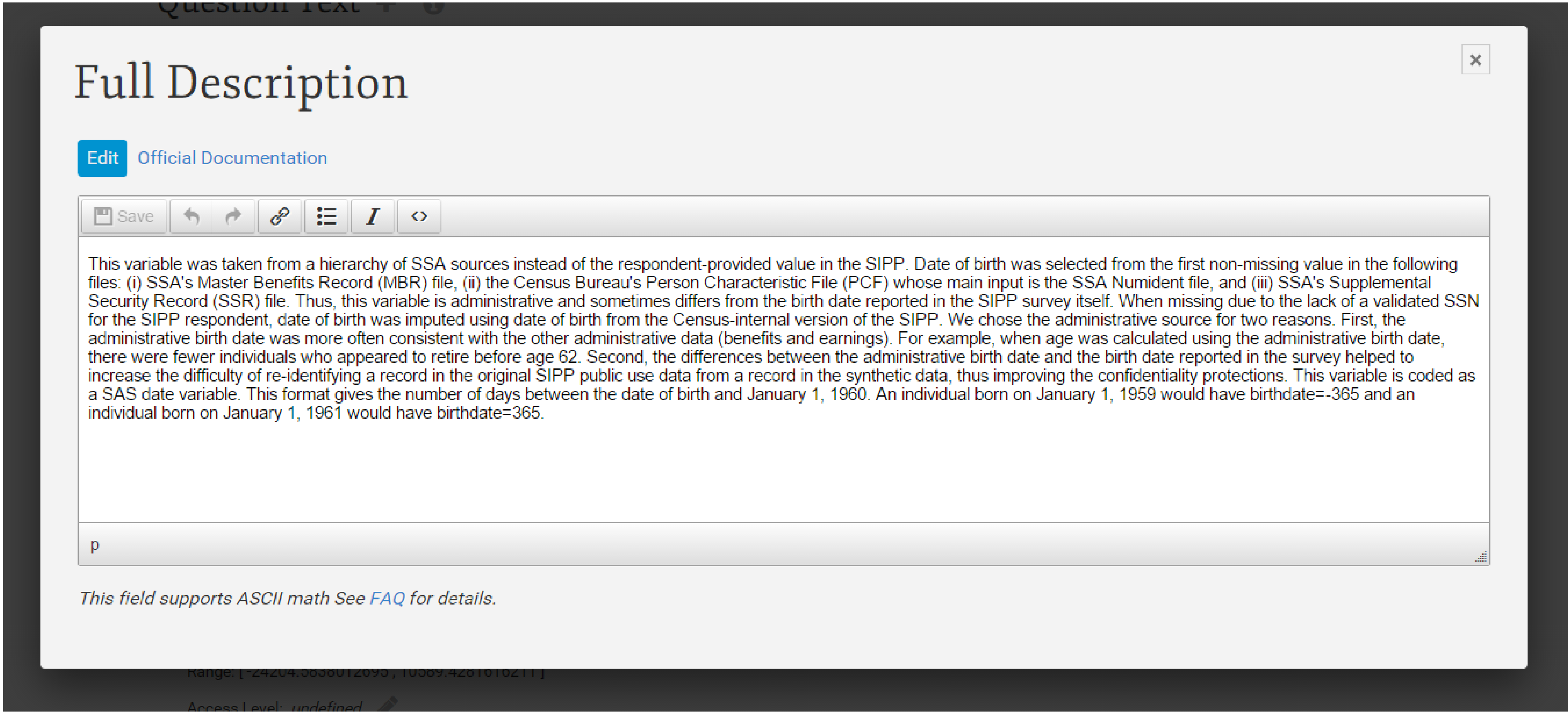
Wiki Branch
(Crowdsource version)

Codebook
rev X

CED^{2}AR User Interface exposes both versions
(with attribution)

Continued Work: Improving merge control

- Merging crowd-sourced content back into official documentation



Full Description

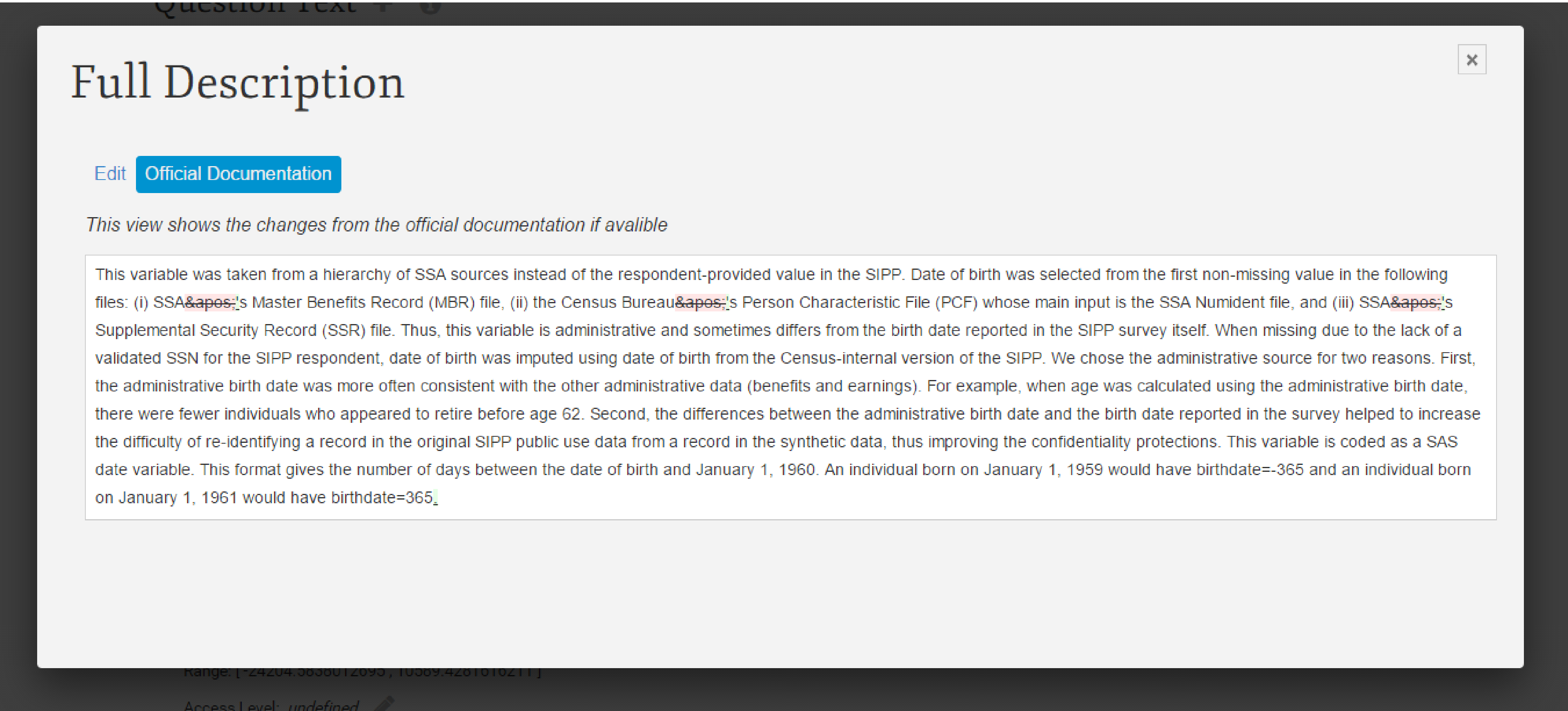
This variable was taken from a hierarchy of SSA sources instead of the respondent-provided value in the SIPP. Date of birth was selected from the first non-missing value in the following files: (i) SSA's Master Benefits Record (MBR) file, (ii) the Census Bureau's Person Characteristic File (PCF) whose main input is the SSA Numident file, and (iii) SSA's Supplemental Security Record (SSR) file. Thus, this variable is administrative and sometimes differs from the birth date reported in the SIPP survey itself. When missing due to the lack of a validated SSN for the SIPP respondent, date of birth was imputed using date of birth from the Census-internal version of the SIPP. We chose the administrative source for two reasons. First, the administrative birth date was more often consistent with the other administrative data (benefits and earnings). For example, when age was calculated using the administrative birth date, there were fewer individuals who appeared to retire before age 62. Second, the differences between the administrative birth date and the birth date reported in the survey helped to increase the difficulty of re-identifying a record in the original SIPP public use data from a record in the synthetic data, thus improving the confidentiality protections. This variable is coded as a SAS date variable. This format gives the number of days between the date of birth and January 1, 1960. An individual born on January 1, 1959 would have birthdate=-365 and an individual born on January 1, 1961 would have birthdate=365.

This field supports ASCII math See FAQ for details.

Access Level: *undefined*

Continued Work: Improving merge control

- Merging crowd-sourced content back into official documentation



Full Description

This view shows the changes from the official documentation if avalible

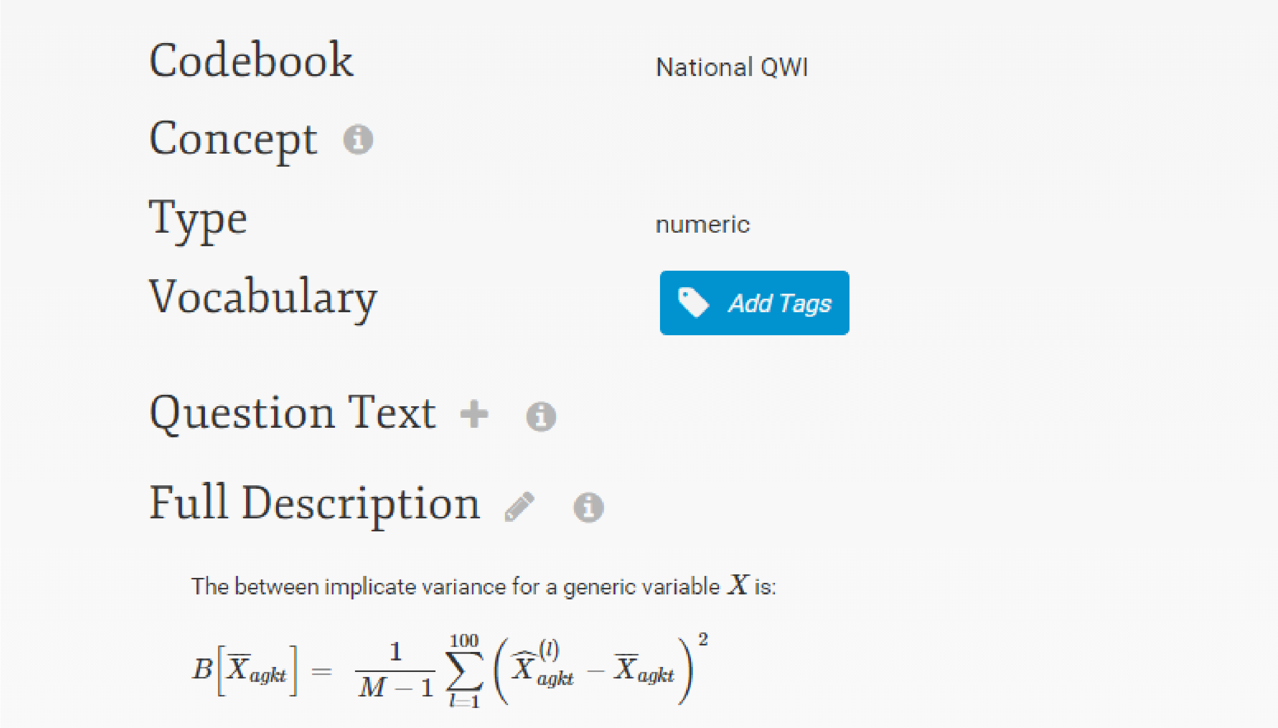
This variable was taken from a hierarchy of SSA sources instead of the respondent-provided value in the SIPP. Date of birth was selected from the first non-missing value in the following files: (i) SSA~~&apos;~~'s Master Benefits Record (MBR) file, (ii) the Census Bureau~~&apos;~~'s Person Characteristic File (PCF) whose main input is the SSA Numident file, and (iii) SSA~~&apos;~~'s Supplemental Security Record (SSR) file. Thus, this variable is administrative and sometimes differs from the birth date reported in the SIPP survey itself. When missing due to the lack of a validated SSN for the SIPP respondent, date of birth was imputed using date of birth from the Census-internal version of the SIPP. We chose the administrative source for two reasons. First, the administrative birth date was more often consistent with the other administrative data (benefits and earnings). For example, when age was calculated using the administrative birth date, there were fewer individuals who appeared to retire before age 62. Second, the differences between the administrative birth date and the birth date reported in the survey helped to increase the difficulty of re-identifying a record in the original SIPP public use data from a record in the synthetic data, thus improving the confidentiality protections. This variable is coded as a SAS date variable. This format gives the number of days between the date of birth and January 1, 1960. An individual born on January 1, 1959 would have birthdate=-365 and an individual born on January 1, 1961 would have birthdate=365.

Range: [-24204.5638012695, 10569.4281616211]

Access Level: *undefined*

Continued Work: Facilitating Editing

- Tagging variables with a controlled vocabulary and a folksonomy



Codebook National QWI

Concept

Type numeric

Vocabulary

Question Text +

Full Description

The between implicate variance for a generic variable X is:

$$B\left[\overline{X}_{agkt}\right] = \frac{1}{M-1}\sum_{l=1}^{100}\left(\widehat{X}_{agkt}^{(l)} - \overline{X}_{agkt}\right)^{2}$$

Thank you!

Questions?

bap63@cornell.edu

ncrn.cornell.edu

github.com/ncrncornell