

Crowdsourcing DDI Development: New Features from the CED²AR Project

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



What is CED²AR?

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



EDDI 2014 “Collaborative editing...”

- 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 CED²AR
 - 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



User Workflow

1. User searches through CED²AR 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

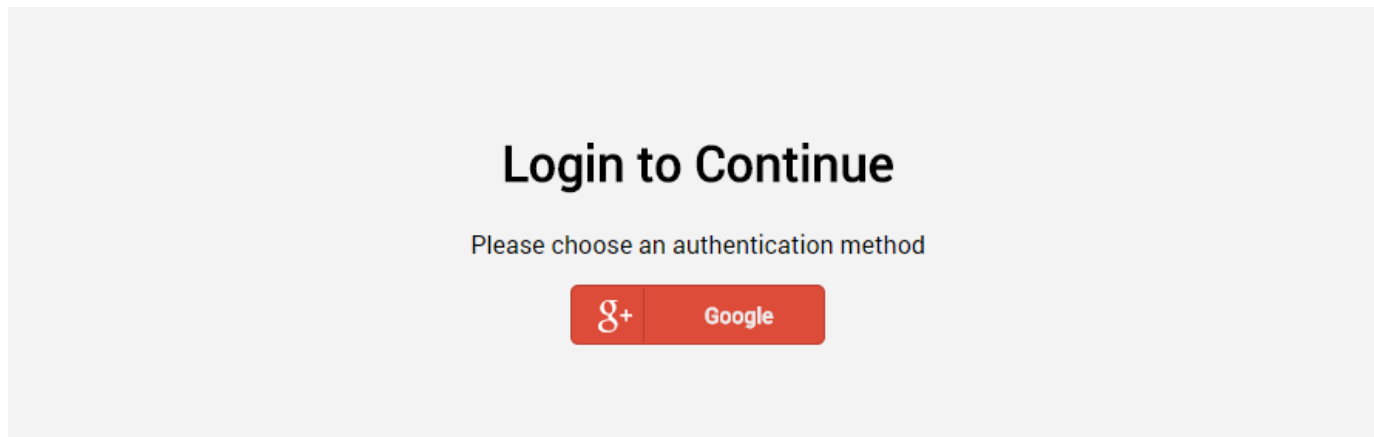
- Expanding the interoperability of DDI

A screenshot of a search engine results page. At the top, a search bar contains the text "ced2ar age" and a blue search button with a magnifying glass icon. Below the search bar, there are navigation tabs for "Web", "Images", "News", "Shopping", "Videos", "More", and "Search tools". The "Web" tab is selected and underlined. Below the tabs, it says "About 11,300 results (0.62 seconds)". The first result is titled "FAQ - CED2AR" with a URL "https://www2.ncrn.cornell.edu/ced2ar-web/docs/faq" and a description: "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 ...". The second result is titled "Age (IPUMSUSA2012) - CED2AR" with a URL "dev.ncrn.cornell.edu/ced2ar-web/codebooks/ipumsusa/v/2012/vars/AGE" and a description: "AGE reports the person's age in years as of the last birthday. ... CED2AR. The Comprehensive Extensible Data Documentation and Access Repository.". The third result is titled "MBR/PHUS Variables Group (SSBV51) - CED2AR" with a URL "https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v51/.../_7/" and a description: "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
- CED²AR handles identity management





Editing

- Automatic validation, and editor for rich content

Data prepared by: Cornell Survey Research Institute

Abstract

Save ↶ ↷ 🔗 ☰ *I* <>

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

p

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



Editing

- Allows for ASCII Math

Full Description ✕

Save ↶ ↷ 🔗 ☰ *I* ⏏

The between implicate variance for a generic variable 'X' is:

``B[̄X_{agkt}] = 1/(M-1)sum_{l=1}^{100}(\hat{X}_{agkt}^{(l)} - ̄X_{agkt})^2``

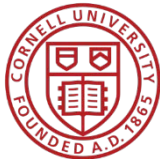
p

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

Full Description

The between implicate variance for a generic variable X is:

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



Editing

- Growing support for additional DDI fields, exposed or not

SIPP Synthetic Beta v6

Producer

View Va
View ob
Last up
Docume
Codebo
Data pr
Data D

Labor Dynamics Institute

<http://www2-urdc.cornell.edu/research/data/sipp-synthetic-beta-file/>

The screenshot shows a web interface for editing a 'Producer' record. The form has a title 'Producer' and a close button. Below the title is a save icon. The form contains four input fields: 'Producer' (containing 'United States Department of Commerce. Bureau of the Census.'), 'Abbreviation' (containing 'Census'), 'Affiliation' (empty), and 'Role' (empty). The background is dark with some text visible on the left side.

Producer	United States Department of Commerce. Bureau of the Census.
Abbreviation	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

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



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)

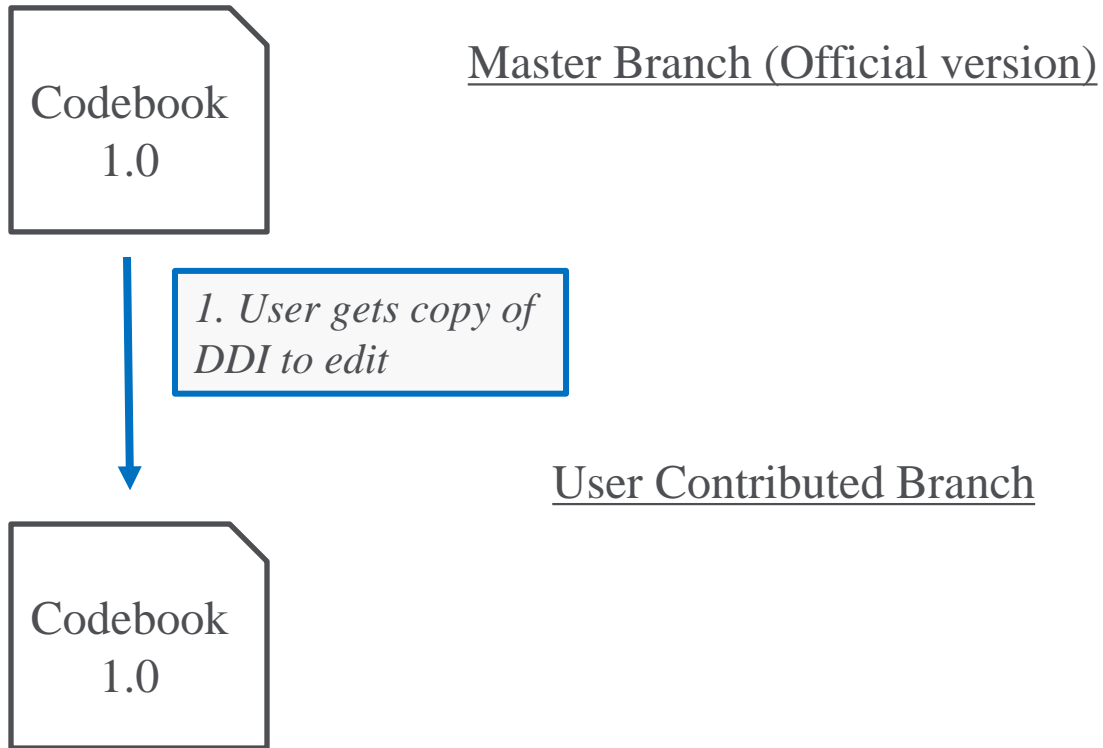
Document Date: June 19th 2014

Codebook prepared by: Cornell NSF Census Research Network

Data prepared by: United States Department of Commerce. Bureau of the Census.

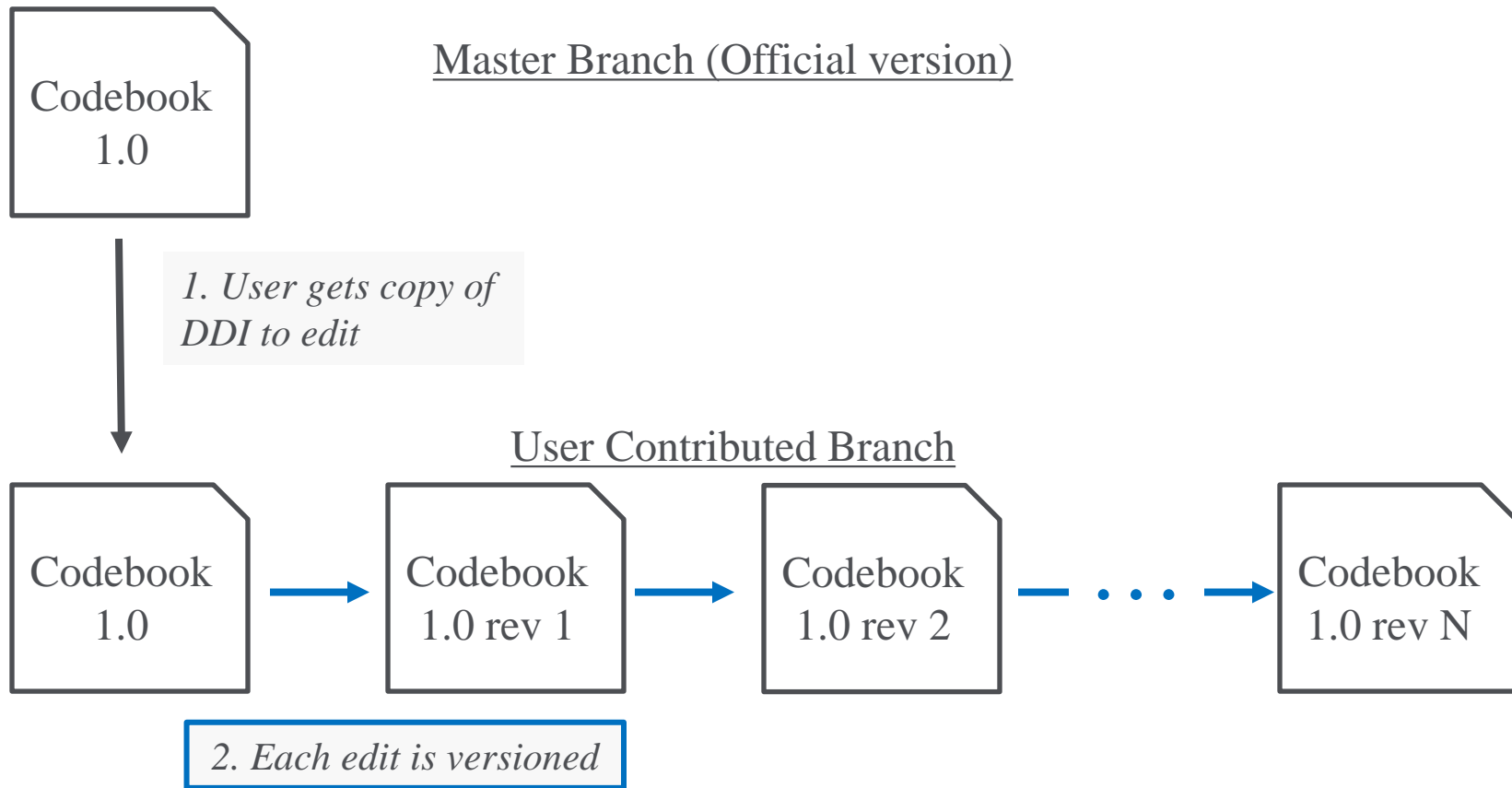


Architecture



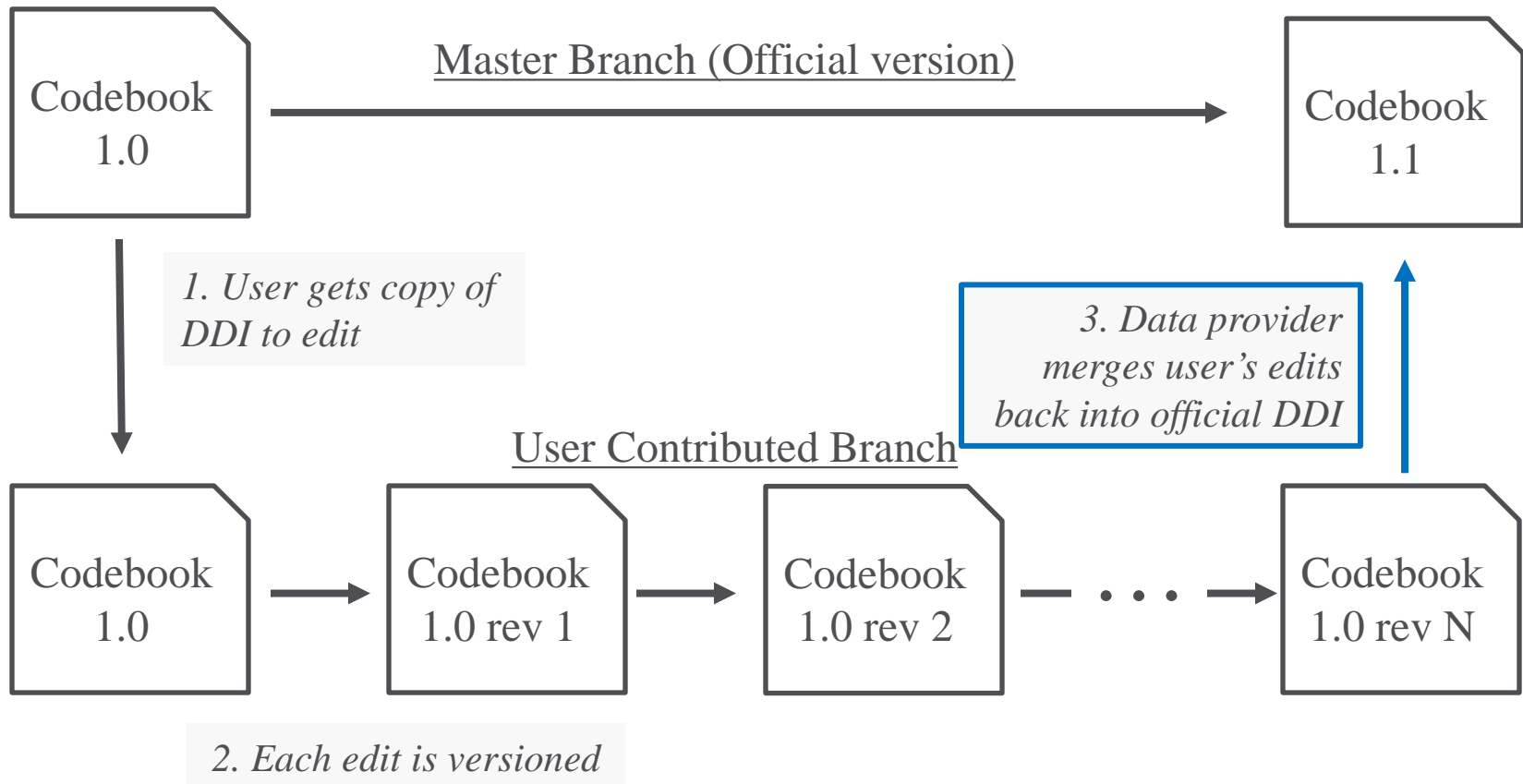


Architecture





Architecture






Architecture

Branches

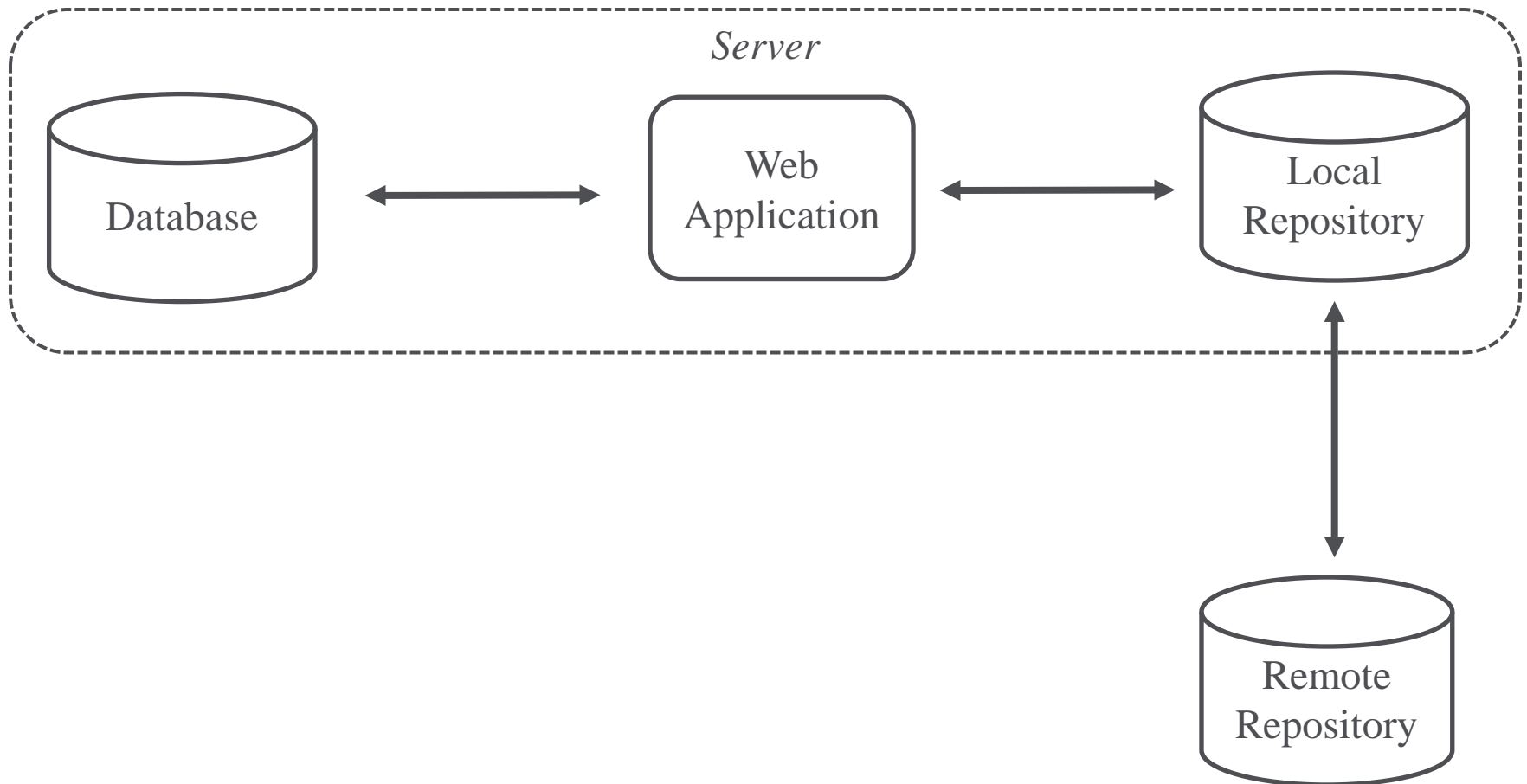
 Create branch

Filters: **Active** Merged

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	



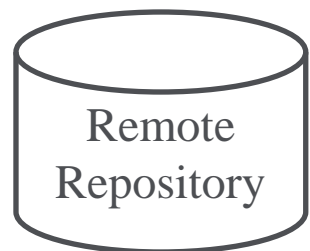
Architecture

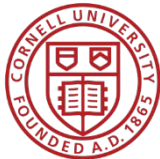




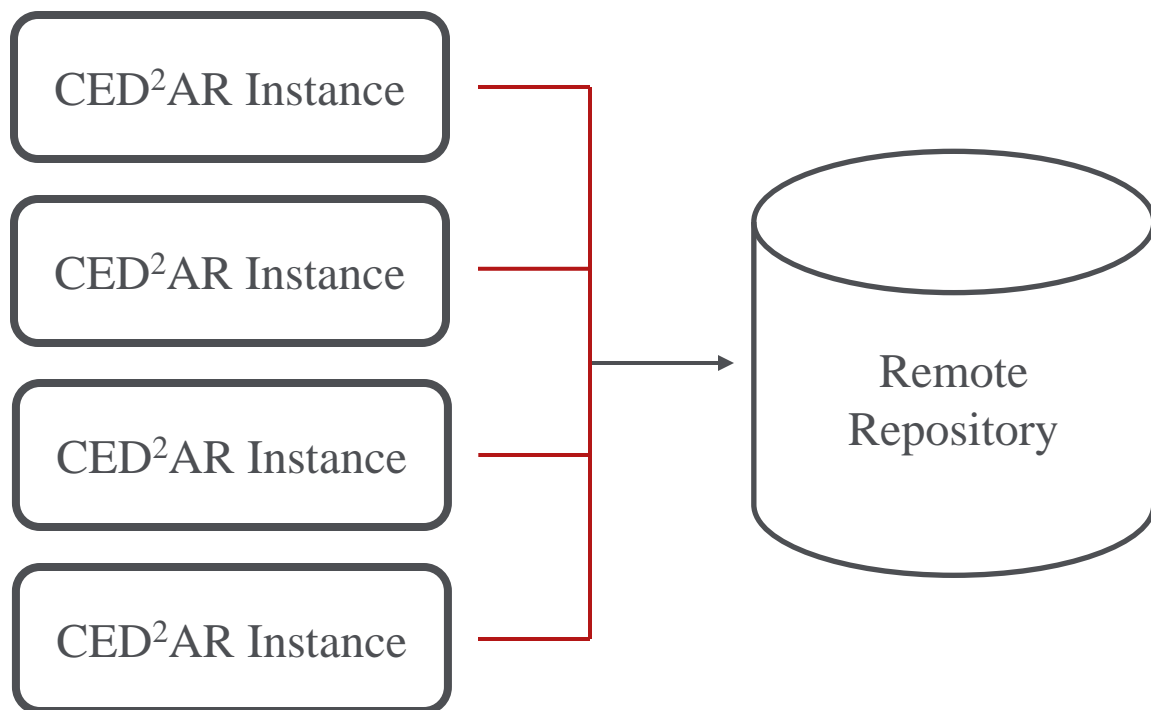
Architecture

CED²AR Instance



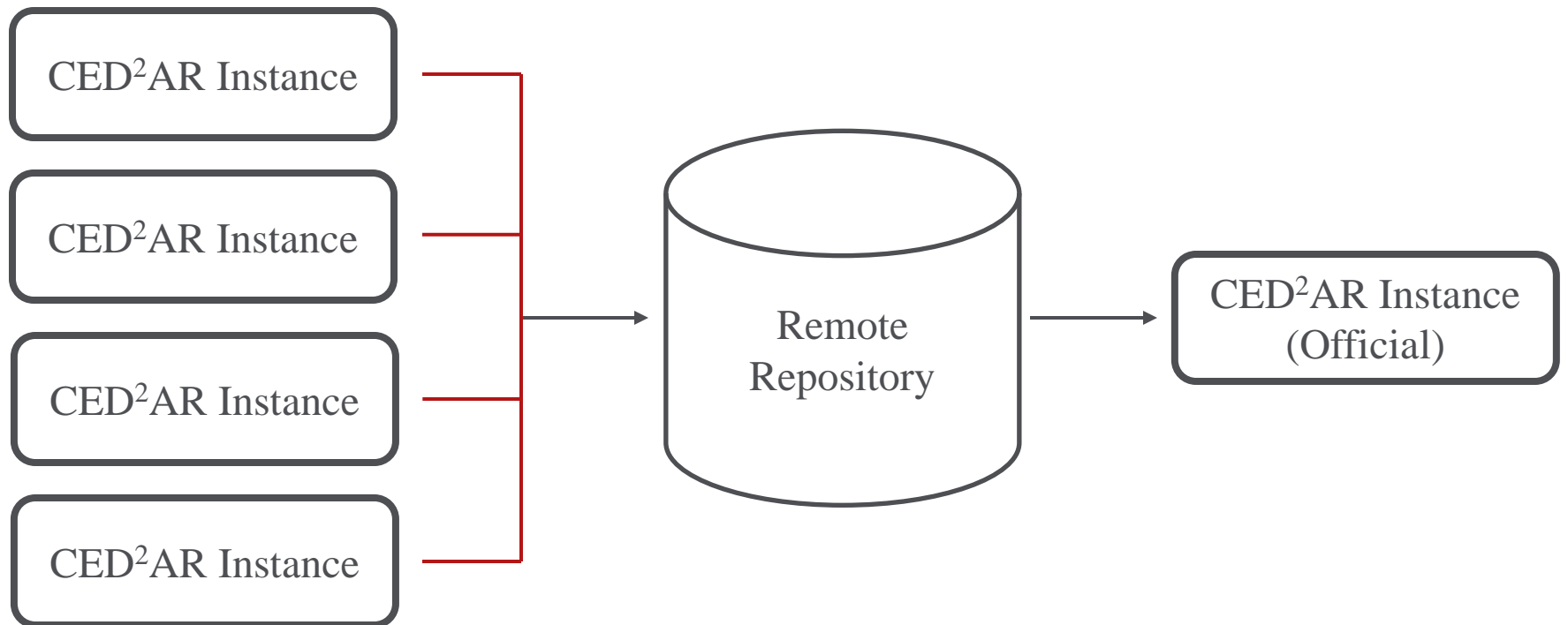


Architecture





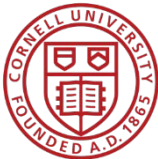
Architecture





Remote Location

- Our implementation uses Bitbucket
- Commit messages describe changes
- Users linked by email address
- Commit hashes are stored on CED²AR
- Remote synchronization is optional



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>
14 14     + <producer abbr="Cornell NCRN Project">Cornell NSF Census Research Network</producer>
15 15     <copyright>Cornell NCRN Project</copyright>
16 16     - <prodDate date="18 June 2014">June 21, 2014</prodDate>
16 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 20     + <software>The Comprehensive Extensible Data Documentation and Access Repository 2.5</software>
21 21     <fundAg abbr="NSF">National Science Foundation (NSF)</fundAg>
22 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 30     - <version date="2014-10-07 09:10:40 (auto-generated)">2014-06-18</version>
31 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>
```



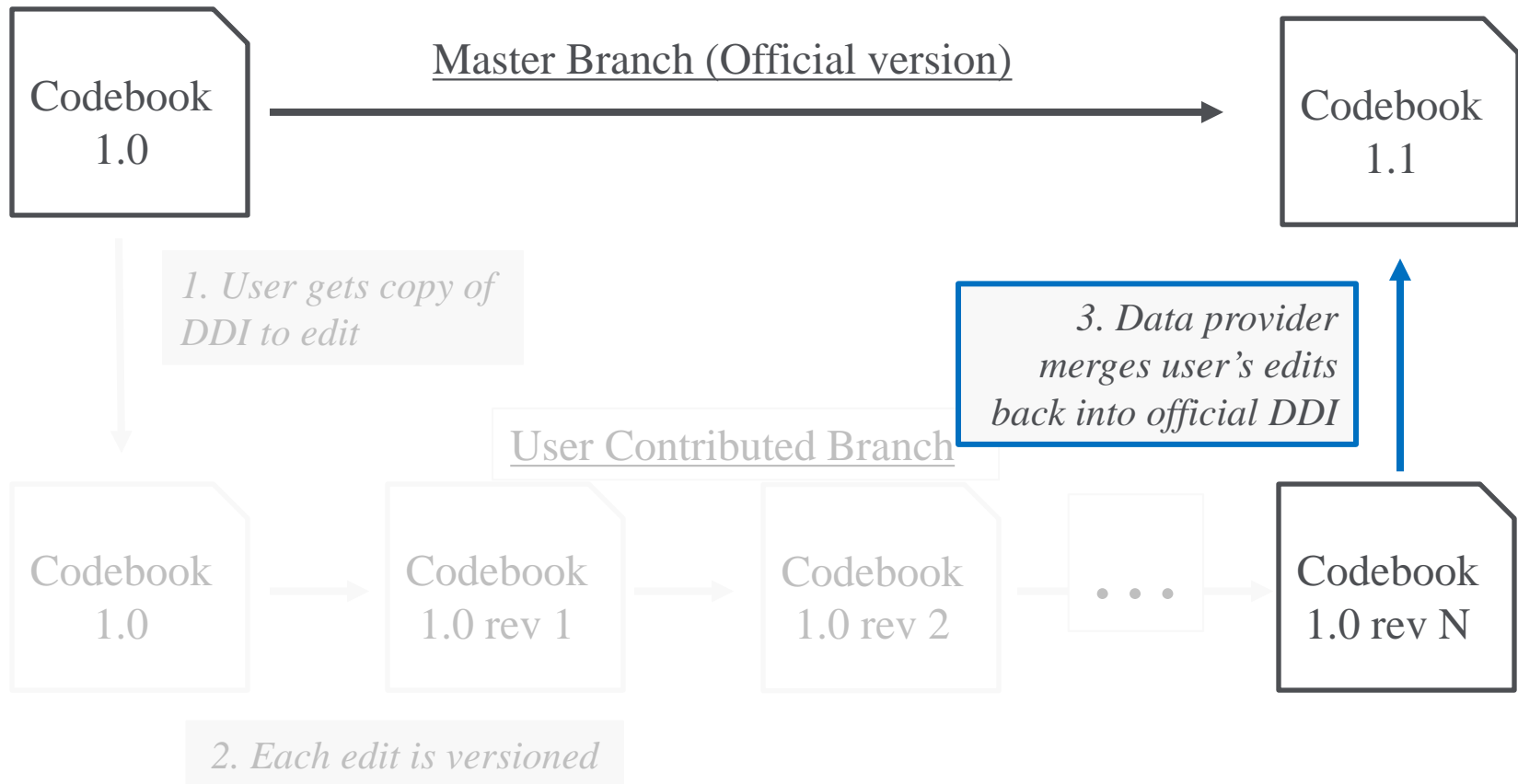
Tracking Changes

Codebook Status

Codebook	Git Status	Last Local Update	BaseX Status
acs.2009.xml	UP_TO_DATE	February 25, 2015 at 11:05 AM: Committing 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: Committing 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: Committing 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}{synlbv2,anonymous,var,act} {synlbv2,anonymous,var,yr}	EXIST_IN_BASEX



Continued Work: Improving merge control





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

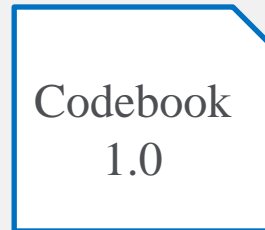


Architecture (alternate)

Master Branch
(Official version)



Wiki Branch
(Community version)



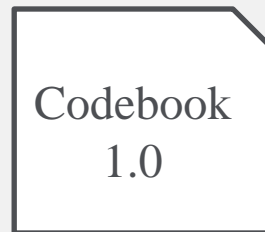


Architecture (alternate)

Master Branch
(Official version)



Wiki Branch
(Community version)



User Branches



*Users pull from wiki branch
into any instance of CED²AR*

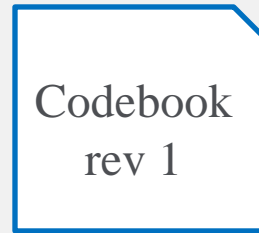
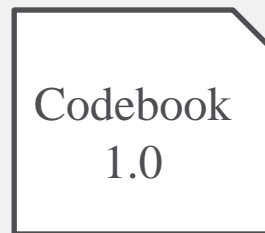


Architecture (alternate)

Master Branch
(Official version)



Wiki Branch
(Community version)



User Branches



Users push back to branch manually

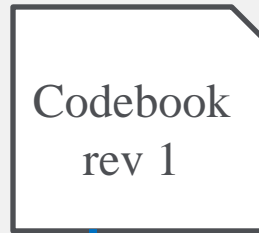
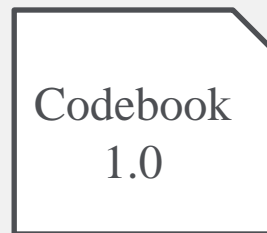


Architecture (alternate)

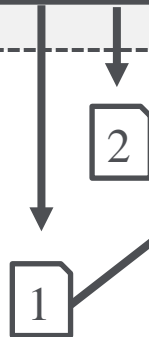
Master Branch
(Official version)



Wiki Branch
(Community version)



User Branches



New users work off most recent revision by default

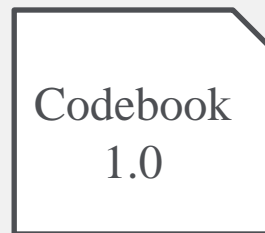


Architecture (alternate)

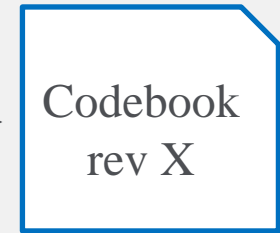
Master Branch
(Official version)



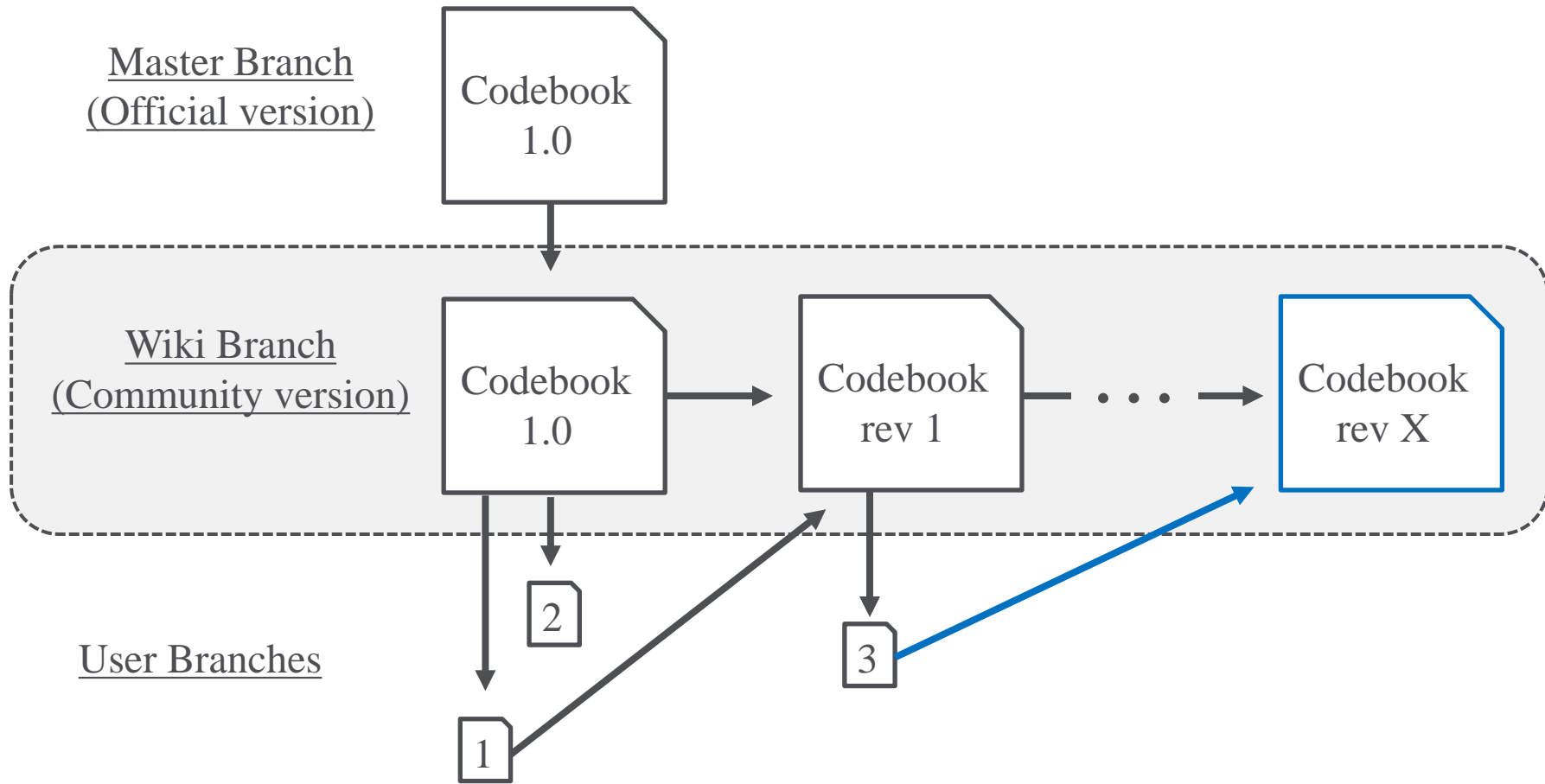
Wiki Branch
(Community version)



...



User Branches



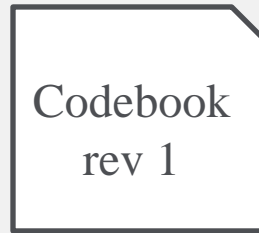


Architecture (alternate)

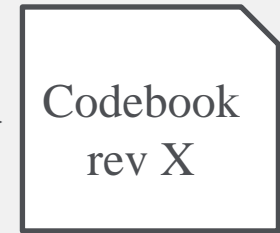
Master Branch
(Official version)



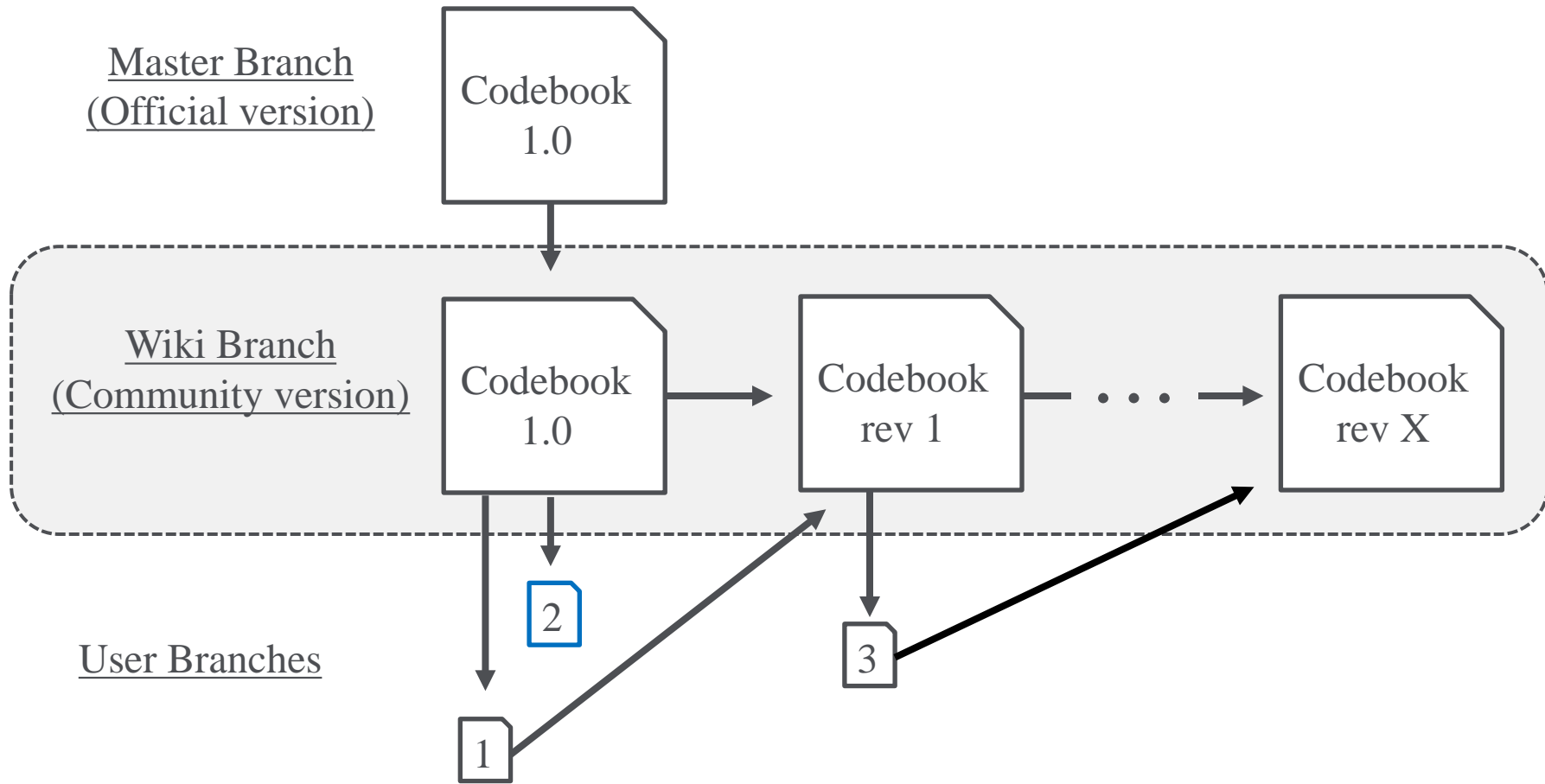
Wiki Branch
(Community version)



...



User Branches





Architecture (alternate)

Master Branch
(Official version)



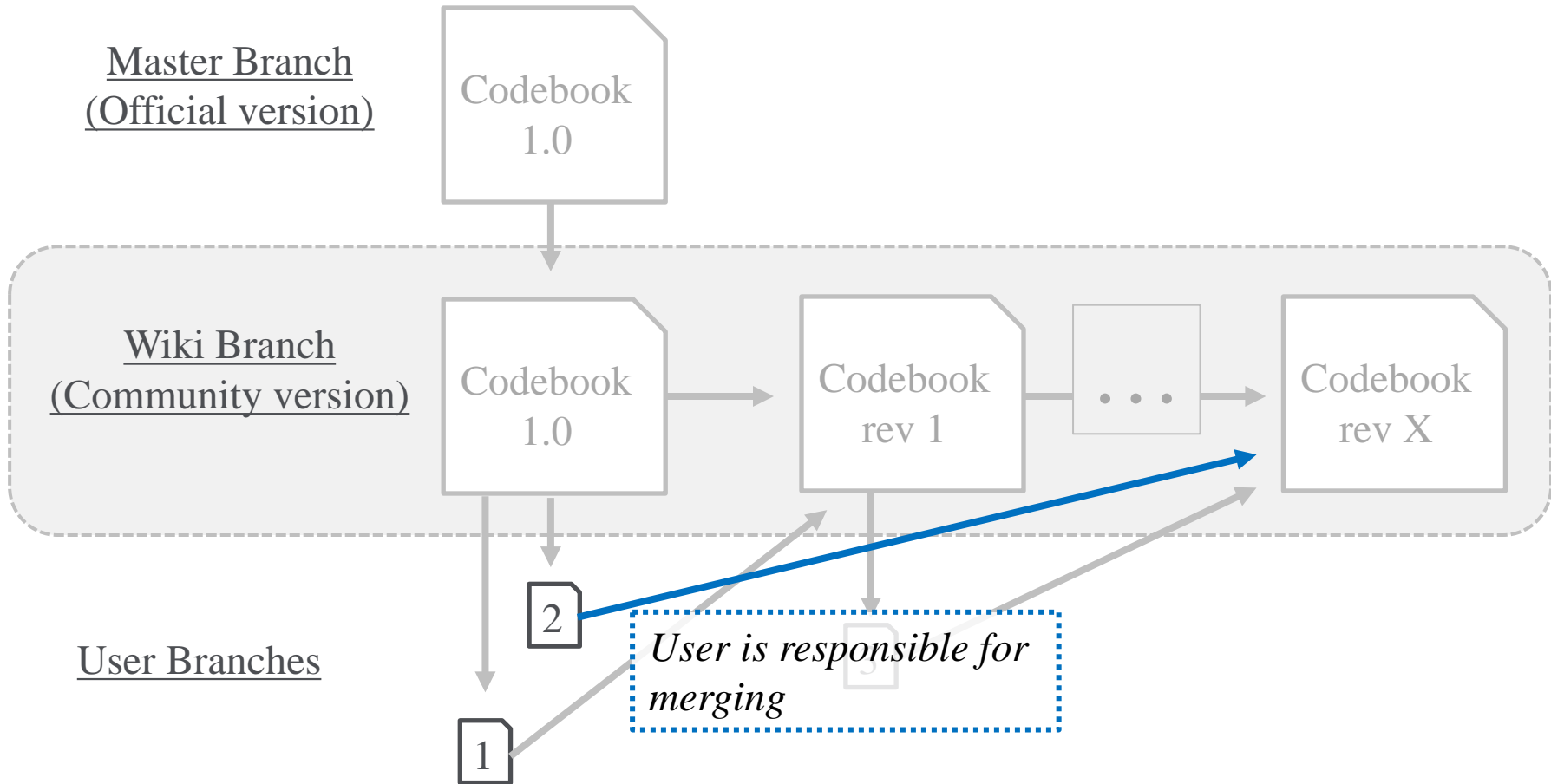
Wiki Branch
(Community version)



User Branches

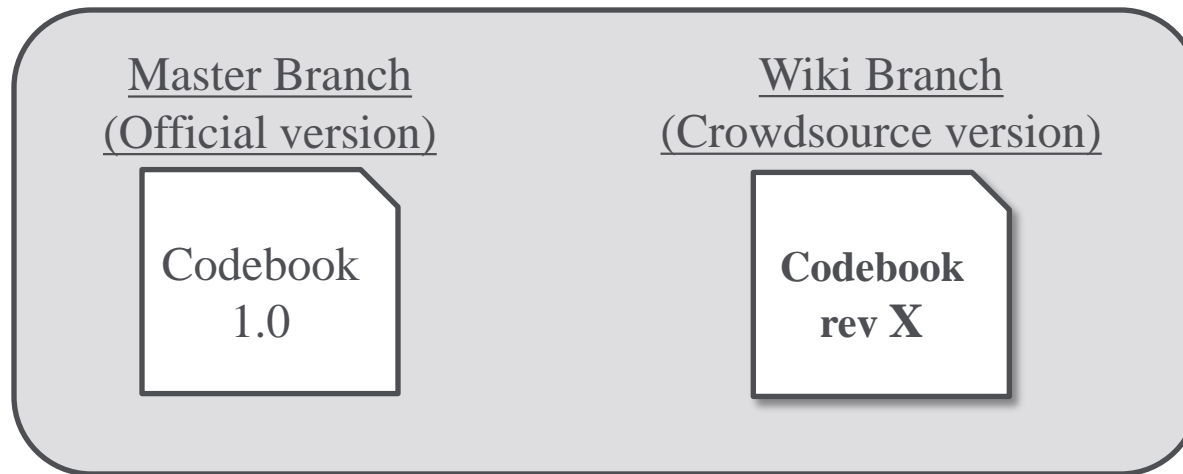


User is responsible for merging





Architecture (alternate)



*CED²AR User
Interface exposes both
versions
(with attribution)*



Continued Work: Improving merge control

- Merging crowd-sourced content back into official documentation

Abstract

The Quarterly Workforce Indicators are local labor market data produced and released every quarter by the United States Census Bureau. Unlike any other local labor market series produced in the U.S. or the rest of the world, the QWI measure employment flows for workers (accession and separations), jobs (creations and destructions) and earnings for demographic subgroups (age and sex), economic industry (NAICS industry groups), and detailed geography (county, Core-Based Statistical Area, and Workforce Investment Area, as well as experimental, unreleased block-level estimates). The current QWI data cover 47 states and about 98% of the private workforce in each of those states.

John Abowd and Lars Vilhuber have used the existing public-use data (and only those public-use data) to construct the first national estimates. The national estimates are an important enhancement to existing series because they include demographic and industry detail for both worker and job flows compiled from data that have been integrated at the micro-level by the Longitudinal Employer-Household Dynamics Program at the Census Bureau. The research paper (see below) compares the new estimates to national data published by the BLS from the Quarterly Census of Employment and Wages and the Business Employment Dynamics series.

p

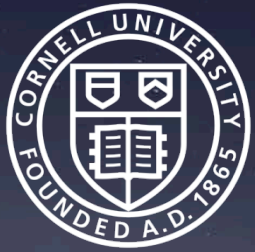
Abstract (User Contributions)

The Quarterly Workforce Indicators are local labor market data produced and released every quarter by the United States Census Bureau. Unlike any other local labor market series produced in the U.S. or the rest of the world, the QWI measure employment flows for workers (accession and separations), jobs (creations and destructions) and earnings for demographic subgroups (age and sex), economic industry (NAICS industry groups), and detailed geography (county, Core-Based Statistical Area, and Workforce Investment Area, as well as experimental, unreleased block-level estimates). **The current QWI data cover 48 states and about 95% of the private workforce in each of those states.**

John Abowd and Lars Vilhuber have used the existing public-use data (and only those public-use data) to construct the first national estimates. The national estimates are an important enhancement to existing series because they include demographic and industry detail for both worker and job flows compiled from data that have been integrated at the micro-level by the Longitudinal Employer-Household Dynamics Program at the Census Bureau. The research paper (see below) compares the new estimates to national data published by the BLS from the Quarterly Census of Employment and Wages and the Business Employment Dynamics series.

p

1 diff found



Thank you!
Questions?

ced2ar-devs-1@cornell.edu

ncrn.cornell.edu

github.com/ncrncornell



Extra slides



Continued Work: Facilitating Editing

- Tagging variables with a controlled vocabulary and a folksonomy

Codebook

National QWI

Concept i

Type

numeric

Vocabulary

 Add Tags

Question Text + i

Full Description  i

The between implicate variance for a generic variable X is:

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



Ingest Workflow

