2018 ADRF Network Research Conference Presentations

ADRF Network Research Conference Presentations

11-2018

Building a repository for record linkage

Susan Hautaniemi Leonard
University of Michigan-Ann Arbor

Abay Israel
University of Michigan-Ann Arbor

Margaret C. Levenstein
University of Michigan-Ann Arbor

Trent Alexander
University of Michigan-Ann Arbor

DOI https://doi.org/10.23889/ijpds.v3i5.1060

Building a repository for record linkage

Abstract
ICPSR is building LinkageLibrary, a repository and community space for researchers involved in linking and combining datasets, as a collaboration between social, statistical, and computer scientists. Unlike surveys or experiments where causal and outcome variables are measured in tandem, it is often necessary when working with organic, non-design data to link to other measures. This makes linkage methodologies particularly important when conducting analyses using administrative data. A common benchmarking repository of linkage methodologies will propel the field to the next level of rigor by facilitating comparison of different algorithms, understanding which types of algorithms work best under different conditions and problem domains, promoting transparency and replicability of research, and encouraging proper citation of methodological contributions and their resulting datasets. It will bring together the diverse scholarly communities (e.g., computer scientists, statisticians, and social, behavioral, economic, and health (SBEH) scientists) who are currently addressing these challenges in disparate ways that do not build on one another's work. Improving linkage methodologies is critical to the production of representative samples, and thus to unbiased estimates of a wide variety of social and economic phenomena. The repository will accelerate the development of new record linkage algorithms and evaluation methods, improve the reproducibility of analyses conducted on integrated data, allow comparisons on same and different data, and move forward the provision of privacy-aware integrated data. The presentation will focus on lessons learned while building the repository and the community, and introduce the LinkageLibrary website.

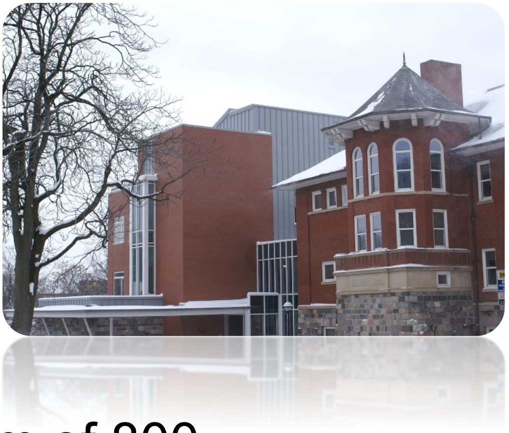
Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1060

LINKAGE LIBRARY

Maintaining datasets to support the data linkage community

linkagelibrary@umich.edu

ICPSR



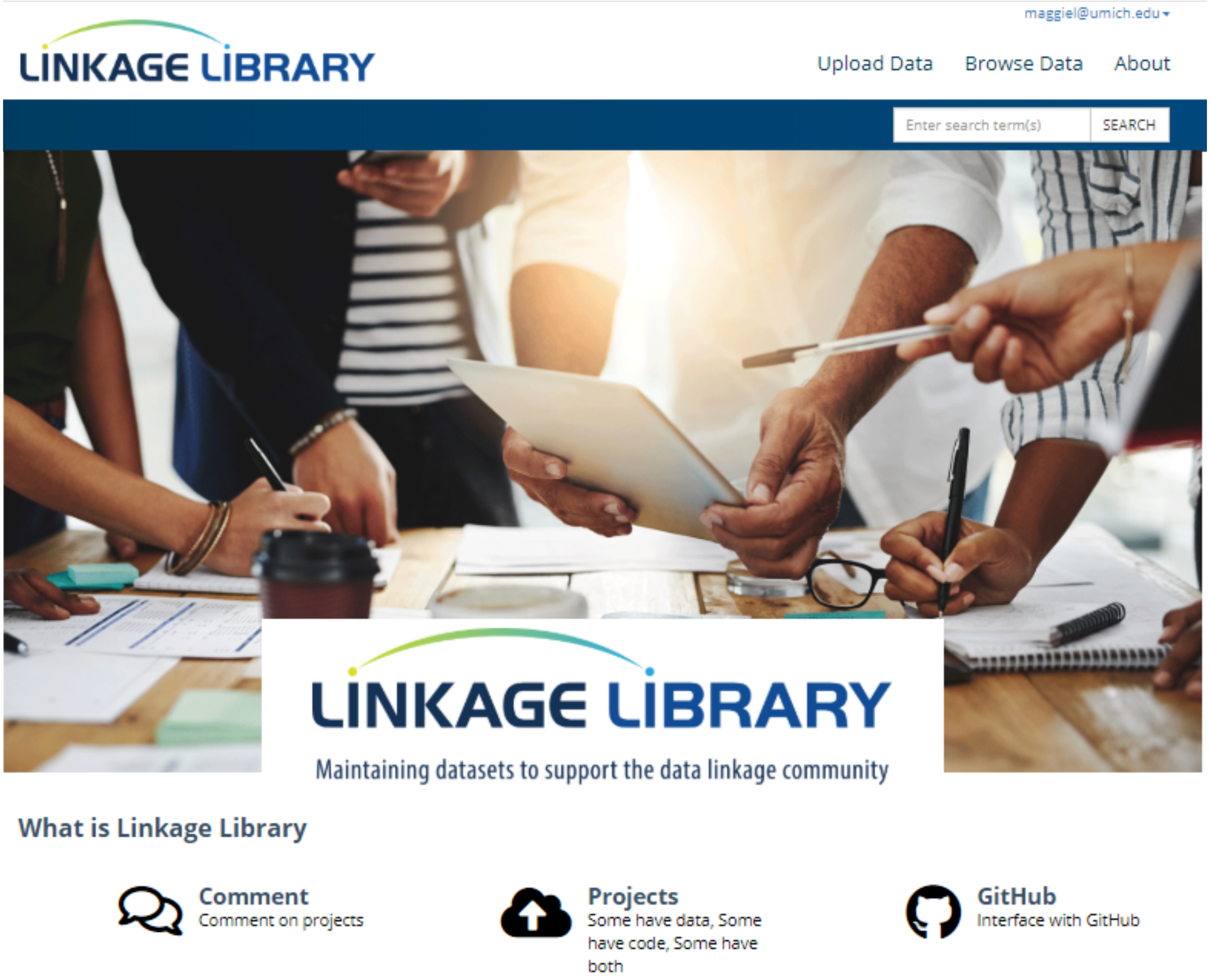
- Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide
- Focus on social and behavioral science data, broadly defined
- Current holdings
  - 10,000 studies, quarter million files, 5 million variables
  - 1500 are *restricted studies*, almost always to protect confidentiality
  - Bibliography of Data-related Literature with 75,000 citations
- Approximately 60,000 active MyData ("shopping cart") accounts
- Thematic collections of data about addiction and HIV, aging, arts and culture, child care and early education, criminal justice, demography, health and medical care, and minorities

Data linkage challenges

- Linked data present challenges for both confidentiality and reproducibility
    - Linkage more accurate with more detailed information
        - Need standards for safe, ethical ways to enhance data with new linkages
    - Linked data easier to re-identify, even after removing unique identifiers
        - Need safe places to analyze linked data
    - Linkage strategies introduce differences in datasets that are often not well documented

LINKAGE LIBRARY

- Encourage researchers to share linked (or linkable) data and linkage strategies
  - Algorithms, code, documents
- Compare approaches across projects, datasets, disciplines
  - Improve linkage practices
  - Improve transparency
  - Commenting – to improve linkage & build community
- Launching in November 2018
  - Looking for beta testers now!
  - Test site



LINKAGE LIBRARY

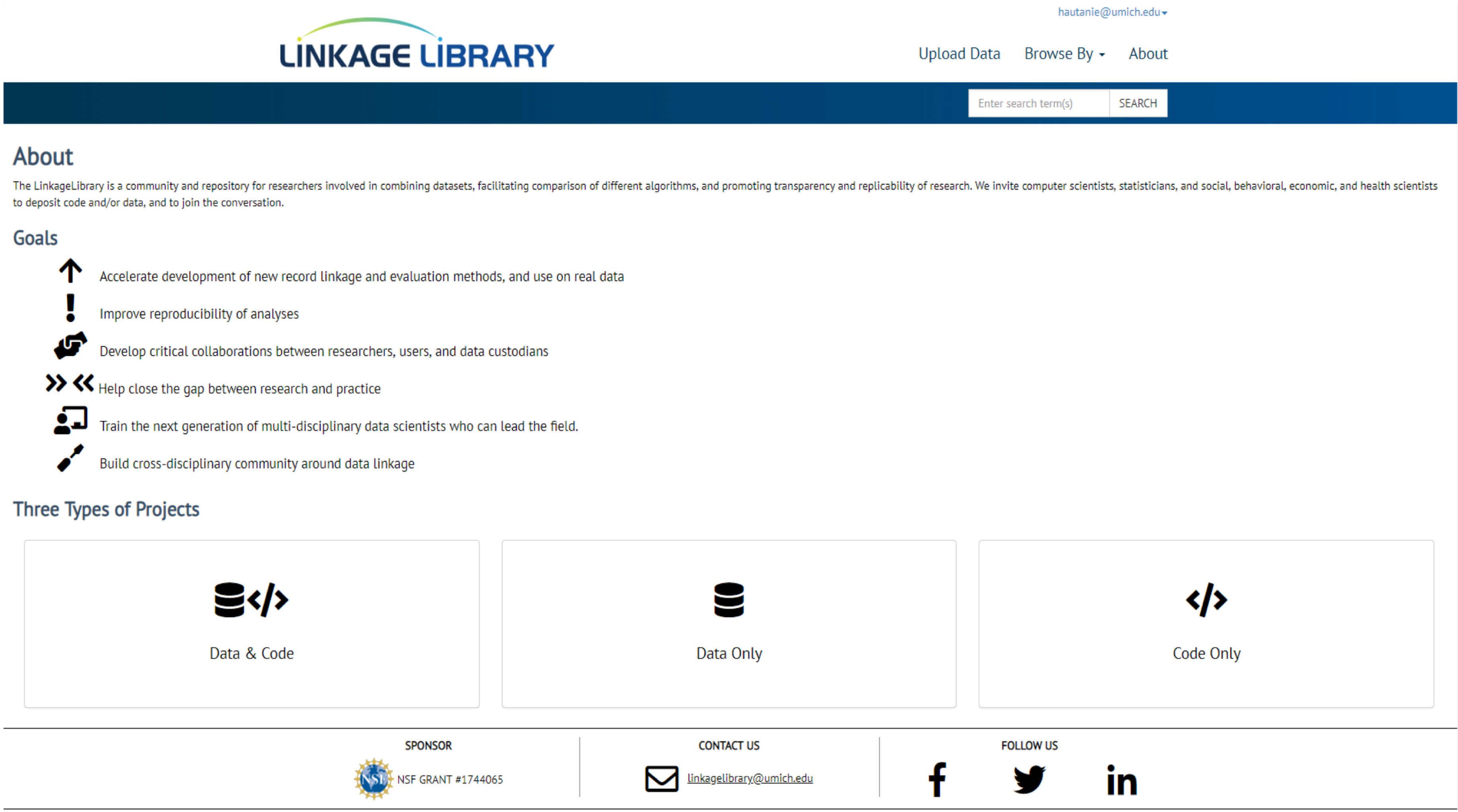
Maintaining datasets to support the data linkage community

What is Linkage Library

Comment
Comment on projects

Projects
Some have data, Some have code, Some have both

GitHub
Interface with GitHub



About

The LinkageLibrary is a community and repository for researchers involved in combining datasets, facilitating comparison of different algorithms, and promoting transparency and replicability of research. We invite computer scientists, statisticians, and social, behavioral, economic, and health scientists to deposit code and/or data, and to join the conversation.

Goals

- Accelerate development of new record linkage and evaluation methods, and use on real data
- Improve reproducibility of analyses
- Develop critical collaborations between researchers, users, and data custodians
- Help close the gap between research and practice
- Train the next generation of multi-disciplinary data scientists who can lead the field.
- Build cross-disciplinary community around data linkage

Three Types of Projects

Data & Code

Data Only

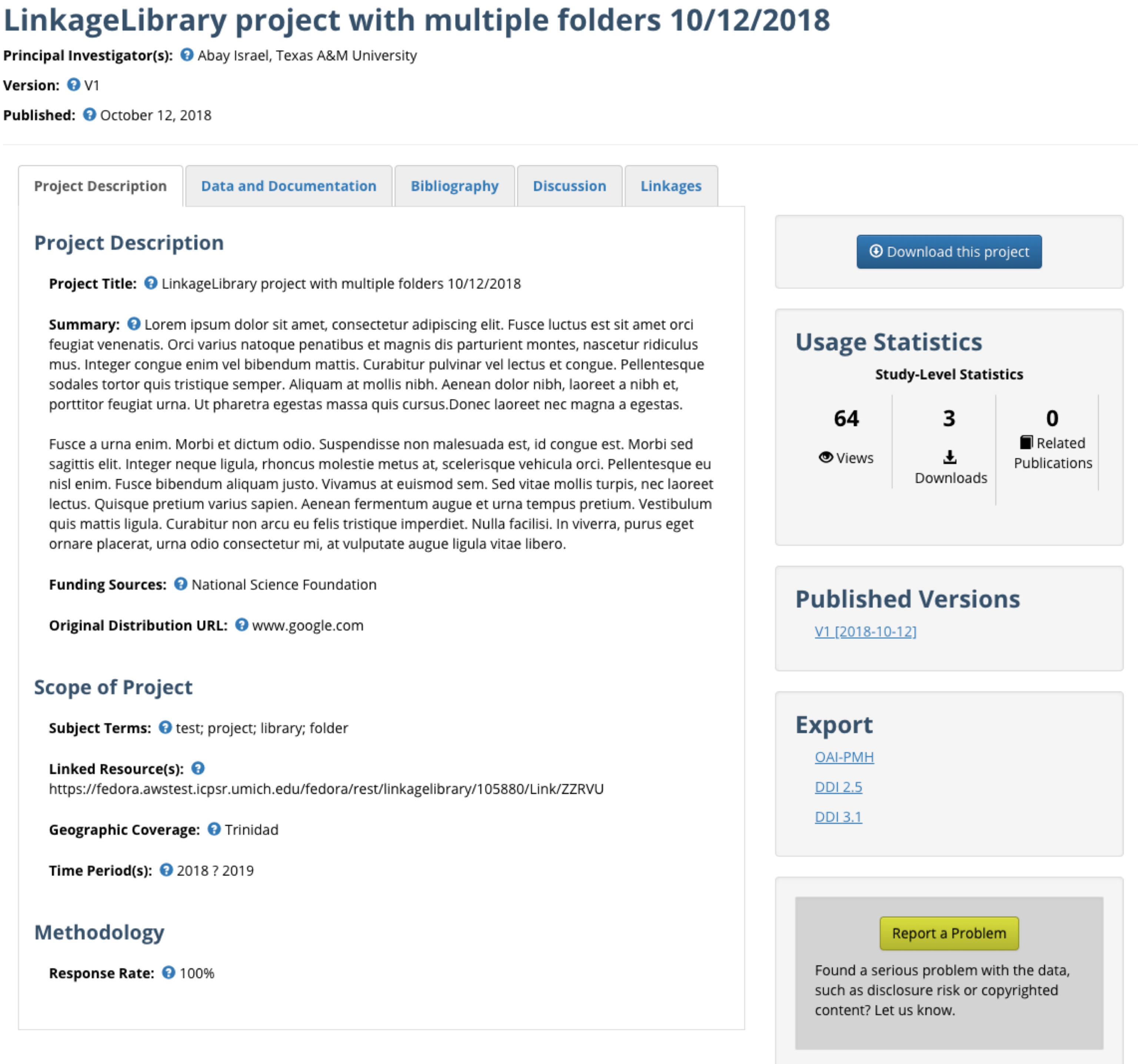
Code Only

SPONSOR

NSF GRANT #1744065

CONTACT US

linkagelibrary@umich.edu



LinkageLibrary project with multiple folders 10/12/2018

Principal Investigator(s): Abay Israel, Texas A&M University

Version: V1

Published: October 12, 2018

Project Description | Data and Documentation | Bibliography | Discussion | Linkages

Project Description

Project Title: LinkageLibrary project with multiple folders 10/12/2018

Summary: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce luctus est sit amet orci feugiat venenatis. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer congue enim vel bibendum mattis. Curabitur pulvinar vel lectus et congue. Pellentesque sodales tortor quis tristique semper. Aliquam at mollis nibh. Aenean dolor nibh, laoreet a nibh et, porttitor feugiat urna. Ut pharetra egestas massa quis cursus.Donec laoreet nec magna a egestas.

Fusce a urna enim. Morbi et dictum odio. Suspendisse non malesuada est, id congue est. Morbi sed sagittis elit. Integer neque ligula, rhoncus molestie metus at, scelerisque vehicula orci. Pellentesque eu nisl enim. Fusce bibendum aliquam justo. Vivamus at euismod sem. Sed vitae mollis turpis, nec laoreet lectus. Quisque pretium varius sapien. Aenean fermentum augue et urna tempus pretium. Vestibulum quis mattis ligula. Curabitur non arcu eu felis tristique imperdiet. Nulla facilisi. In viverra, purus eget ornare placerat, urna odio consectetur mi, at vulputate augue ligula vitae libero.

Funding Sources: National Science Foundation

Original Distribution URL: www.google.com

Scope of Project

Subject Terms: test; project; library; folder

Linked Resource(s):
https://fedora.awstest.icpsr.umich.edu/fedora/rest/linkagelibrary/105880/Link/ZZRVU

Geographic Coverage: Trinidad

Time Period(s): 2018 ? 2019

Methodology

Response Rate: 100%

Usage Statistics

Study-Level Statistics

64	3	0
Views	Downloads	Related Publications

Published Versions

V1 [2018-10-12]

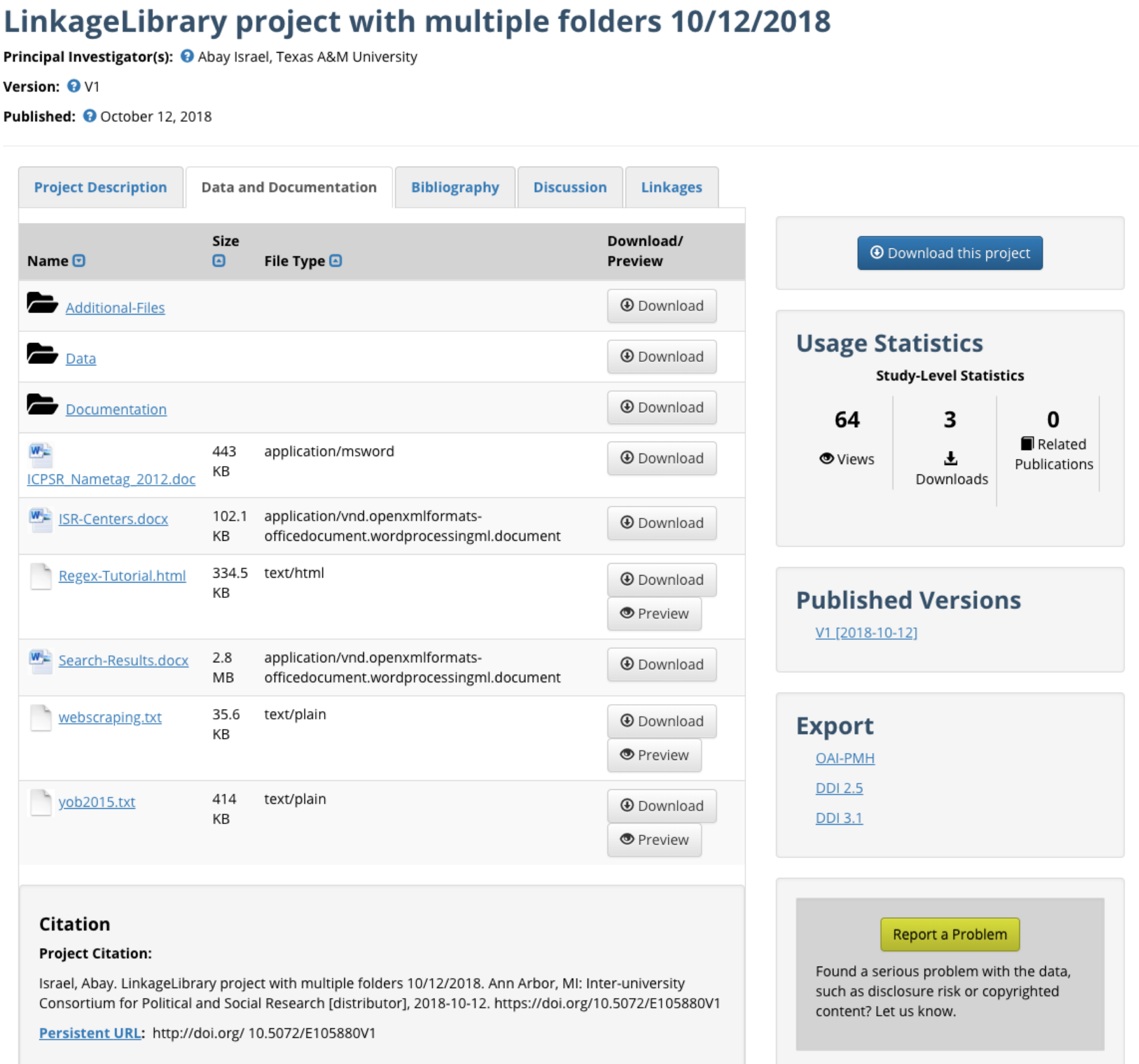
Export

OAI-PMH

DDI 2.5

DDI 3.1

Found a serious problem with the data, such as disclosure risk or copyrighted content? Let us know.



LinkageLibrary project with multiple folders 10/12/2018

Principal Investigator(s): Abay Israel, Texas A&M University

Version: V1

Published: October 12, 2018

Name	Size	File Type	Download/ Preview
Additional-Files			Download
Data			Download
Documentation			Download
ICPSR_Nametag_2012.doc	443 KB	application/msword	Download
ISR-Centers.docx	102.1 KB	application/vnd.openxmlformats-officedocument.wordprocessingml.document	Download
Regex-Tutorial.html	334.5 KB	text/html	Download Preview
Search-Results.docx	2.8 MB	application/vnd.openxmlformats-officedocument.wordprocessingml.document	Download
webscraping.txt	35.6 KB	text/plain	Download Preview
yob2015.txt	414 KB	text/plain	Download Preview

Citation

Project Citation:

Israel, Abay. LinkageLibrary project with multiple folders 10/12/2018. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-10-12. https://doi.org/10.5072/E105880V1

Persistent URL: http://doi.org/ 10.5072/E105880V1

Usage Statistics

Study-Level Statistics

64	3	0
Views	Downloads	Related Publications

Published Versions

V1 [2018-10-12]

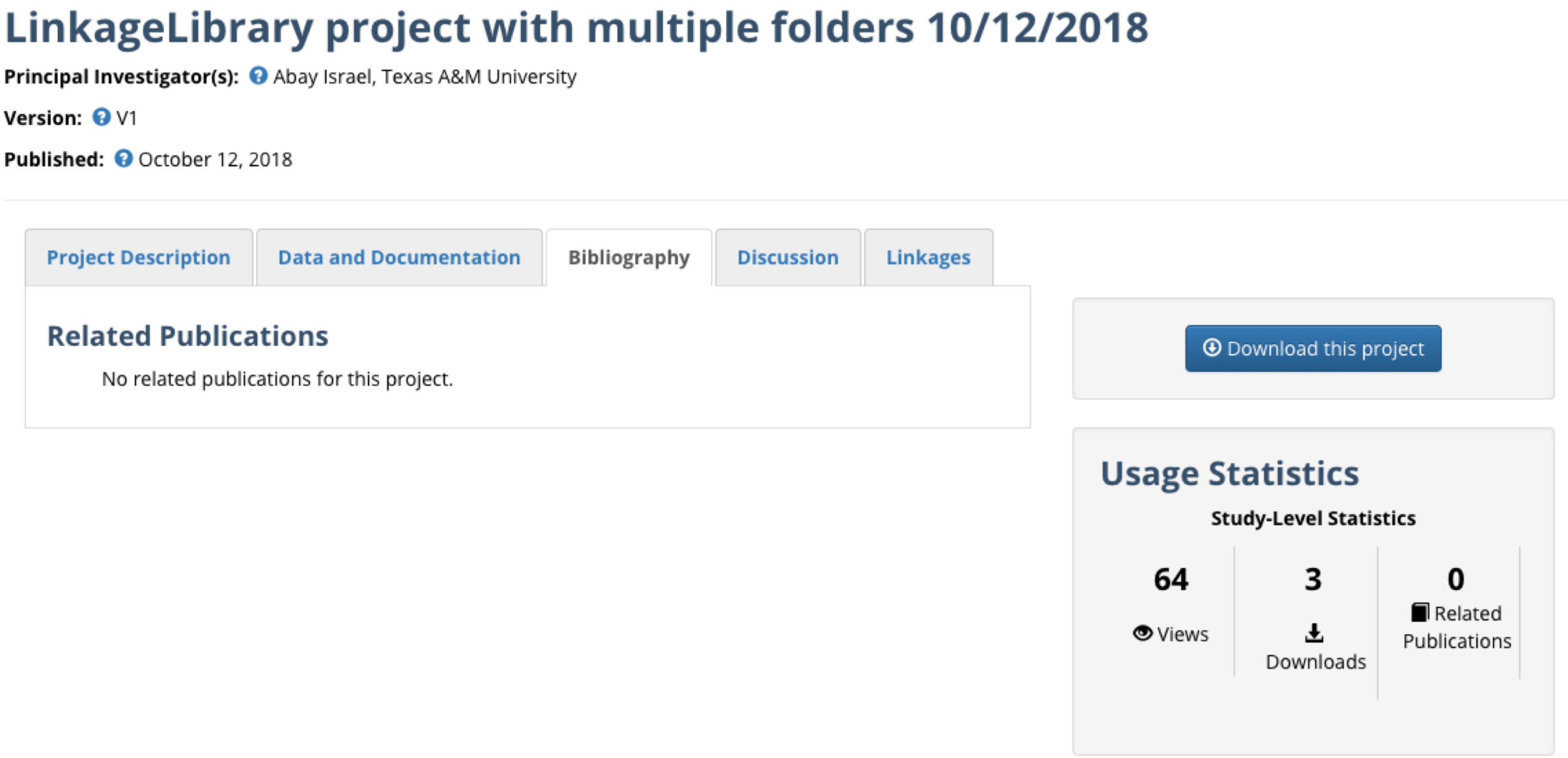
Export

OAI-PMH

DDI 2.5

DDI 3.1

Found a serious problem with the data, such as disclosure risk or copyrighted content? Let us know.



LinkageLibrary project with multiple folders 10/12/2018

Principal Investigator(s): Abay Israel, Texas A&M University

Version: V1

Published: October 12, 2018

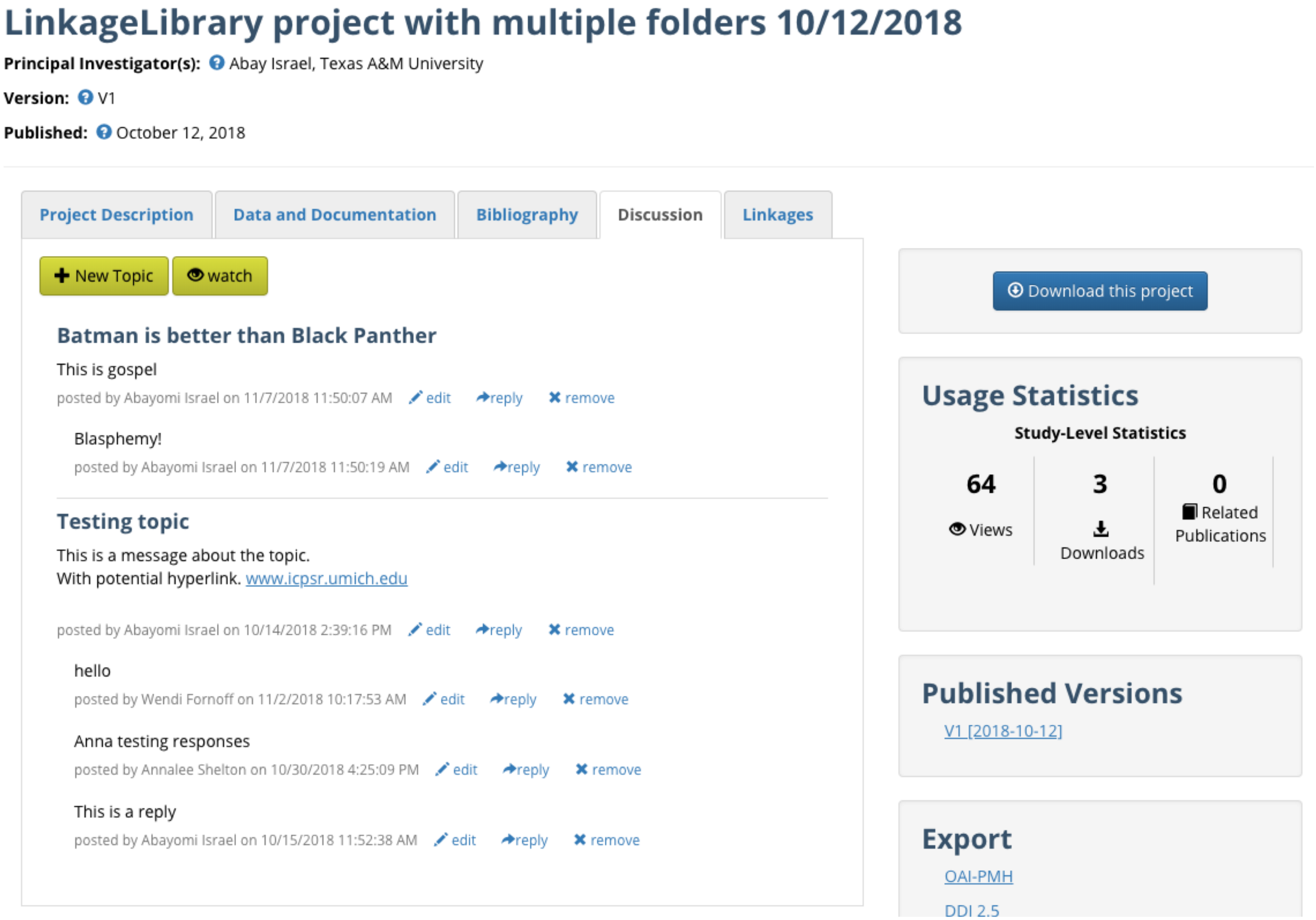
Related Publications

No related publications for this project.

Usage Statistics

Study-Level Statistics

64	3	0
Views	Downloads	Related Publications



LinkageLibrary project with multiple folders 10/12/2018

Principal Investigator(s): Abay Israel, Texas A&M University

Version: V1

Published: October 12, 2018

Batman is better than Black Panther

This is gospel

posted by Abayomi Israel on 11/7/2018 11:50:07 AM edit reply remove

Blasphemy!

posted by Abayomi Israel on 11/7/2018 11:50:19 AM edit reply remove

Testing topic

This is a message about the topic.
With potential hyperlink. www.icpsr.umich.edu

posted by Abayomi Israel on 10/14/2018 2:39:16 PM edit reply remove

hello

posted by Wendi Fornoff on 11/2/2018 10:17:53 AM edit reply remove

Anna testing responses

posted by Annalee Shelton on 10/30/2018 4:25:09 PM edit reply remove

This is a reply

posted by Abayomi Israel on 10/15/2018 11:52:38 AM edit reply remove

Usage Statistics

Study-Level Statistics

64	3	0
Views	Downloads	Related Publications

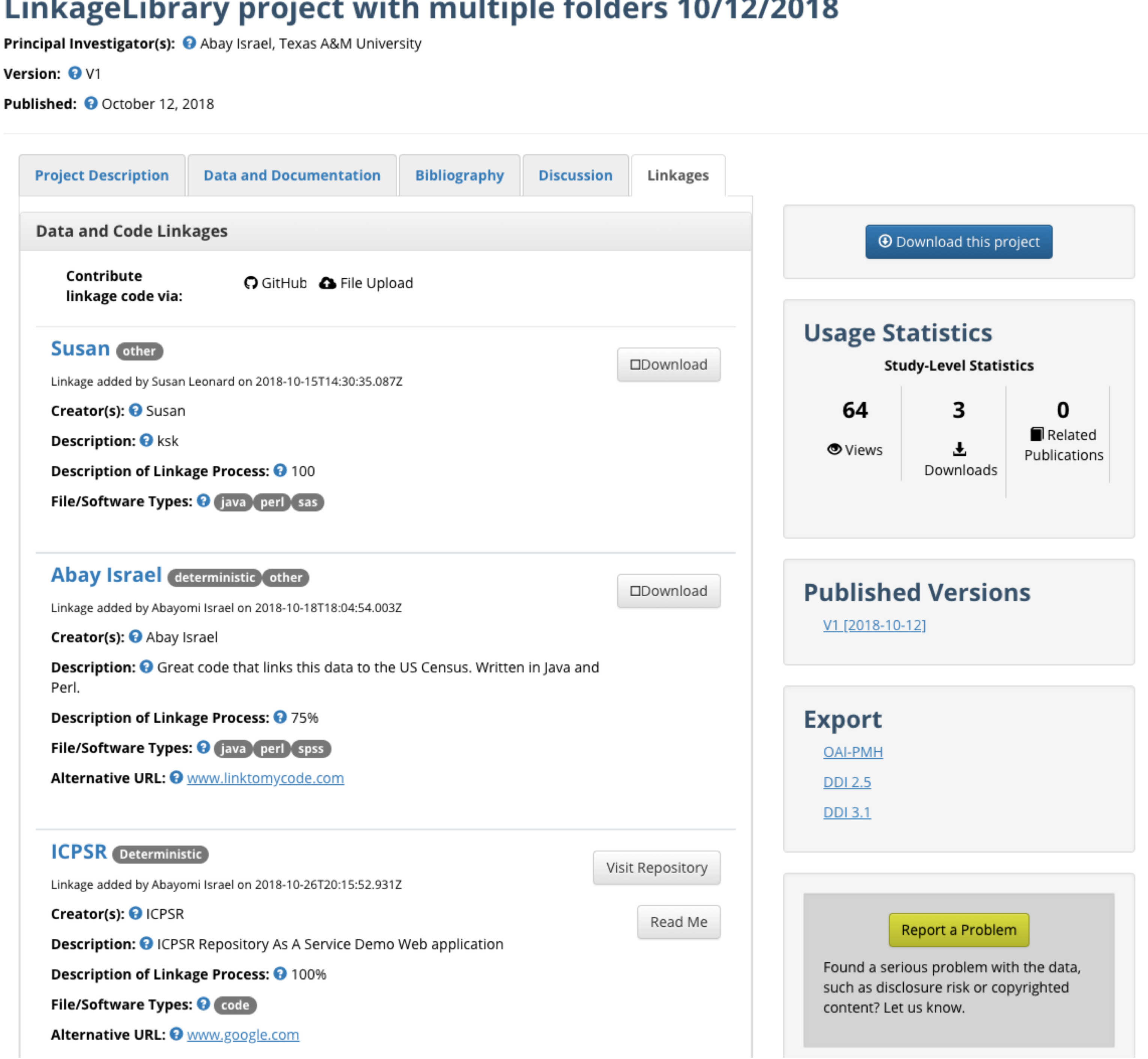
Published Versions

V1 [2018-10-12]

Export

OAI-PMH

DDI 2.5



LinkageLibrary project with multiple folders 10/12/2018

Principal Investigator(s): Abay Israel, Texas A&M University

Version: V1

Published: October 12, 2018

Data and Code Linkages

Contribute linkage code via: GitHub File Upload

Susan other

Linkage added by Susan Leonard on 2018-10-15T14:30:35.087Z

Creator(s): Susan

Description: ksk

Description of Linkage Process: 100

File/Software Types: java perl sas

Abay Israel deterministic other

Linkage added by Abayomi Israel on 2018-10-18T18:04:54.003Z

Creator(s): Abay Israel

Description: Great code that links this data to the US Census. Written in Java and Perl.

Description of Linkage Process: 75%

File/Software Types: java perl spss

Alternative URL: www.linktomycode.com

ICPSR Deterministic

Linkage added by Abayomi Israel on 2018-10-26T20:15:52.931Z

Creator(s): ICPSR

Description: ICPSR Repository As A Service Demo Web application

Description of Linkage Process: 100%

File/Software Types: code

Alternative URL: www.google.com

Usage Statistics

Study-Level Statistics

64	3	0
Views	Downloads	Related Publications

Published Versions

V1 [2018-10-12]

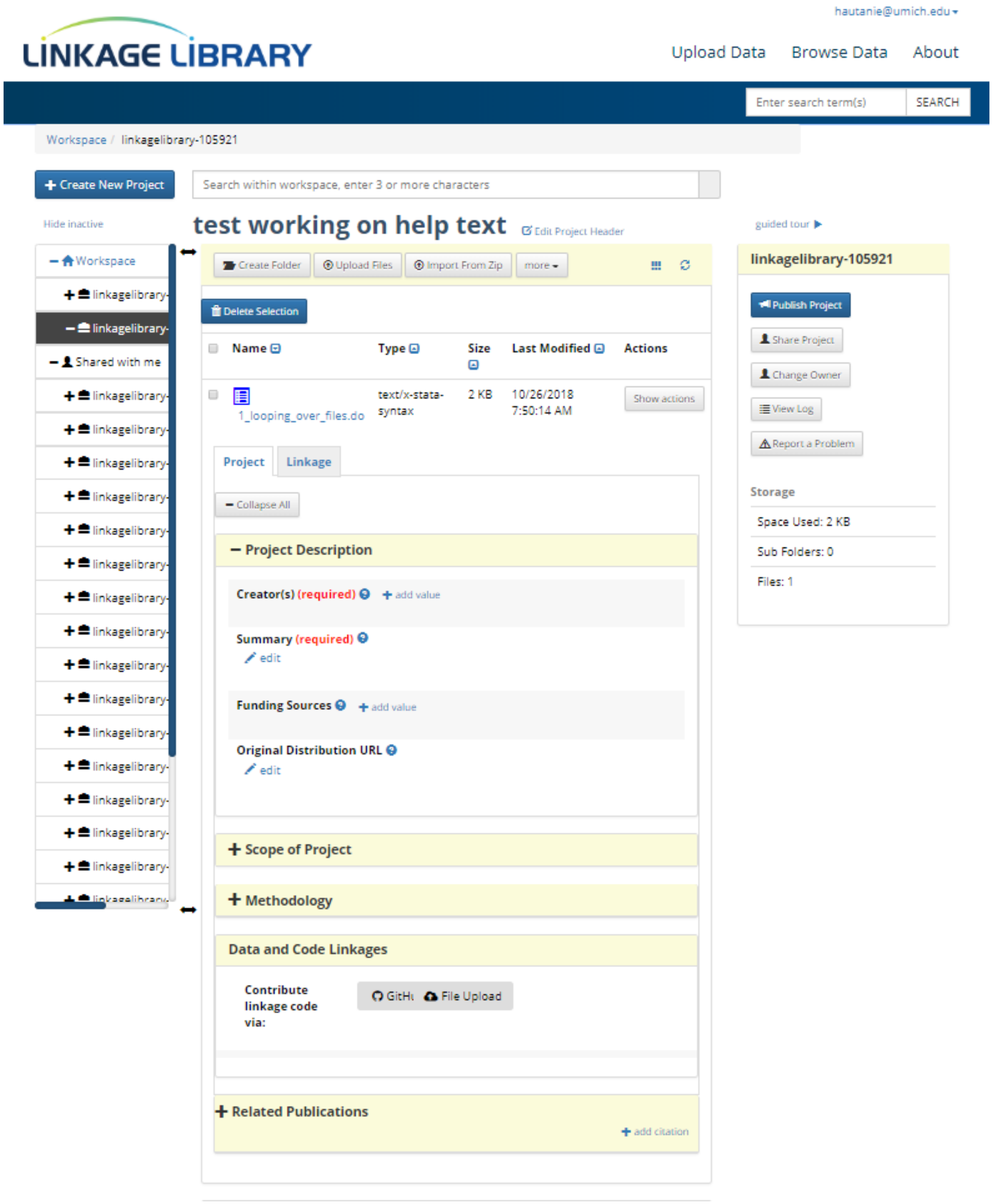
Export

OAI-PMH

DDI 2.5

DDI 3.1

Found a serious problem with the data, such as disclosure risk or copyrighted content? Let us know.



hautanie@umich.edu ▾

LINKAGE LIBRARY

Upload Data Browse Data About

Enter search term(s) SEARCH

Workspace / linkagelibrary-105921

+ Create New Project

Search within workspace, enter 3 or more characters

Hide inactive

− Workspace
+ linkagelibrary-
− linkagelibrary-
− Shared with me
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-
+ linkagelibrary-

test working on help text

Edit Project Header

guided tour ▶

Create Folder | Upload Files | Import From Zip | more ▾

Delete Selection

□	Name	Type	Size	Last Modified	Actions
□	1_looping_over_files.do	text/x-stata-syntax	2 KB	10/26/2018 7:50:14 AM	Show actions

Project | Linkage

− Collapse All

− Project Description

Creator(s) (required) + add value

Summary (required)
edit

Funding Sources + add value

Original Distribution URL
edit

+ Scope of Project

+ Methodology

Data and Code Linkages

Contribute linkage code via: GitHı File Upload

+ Related Publications

+ add citation

linkagelibrary-105921

Publish Project

Share Project

Change Owner

View Log

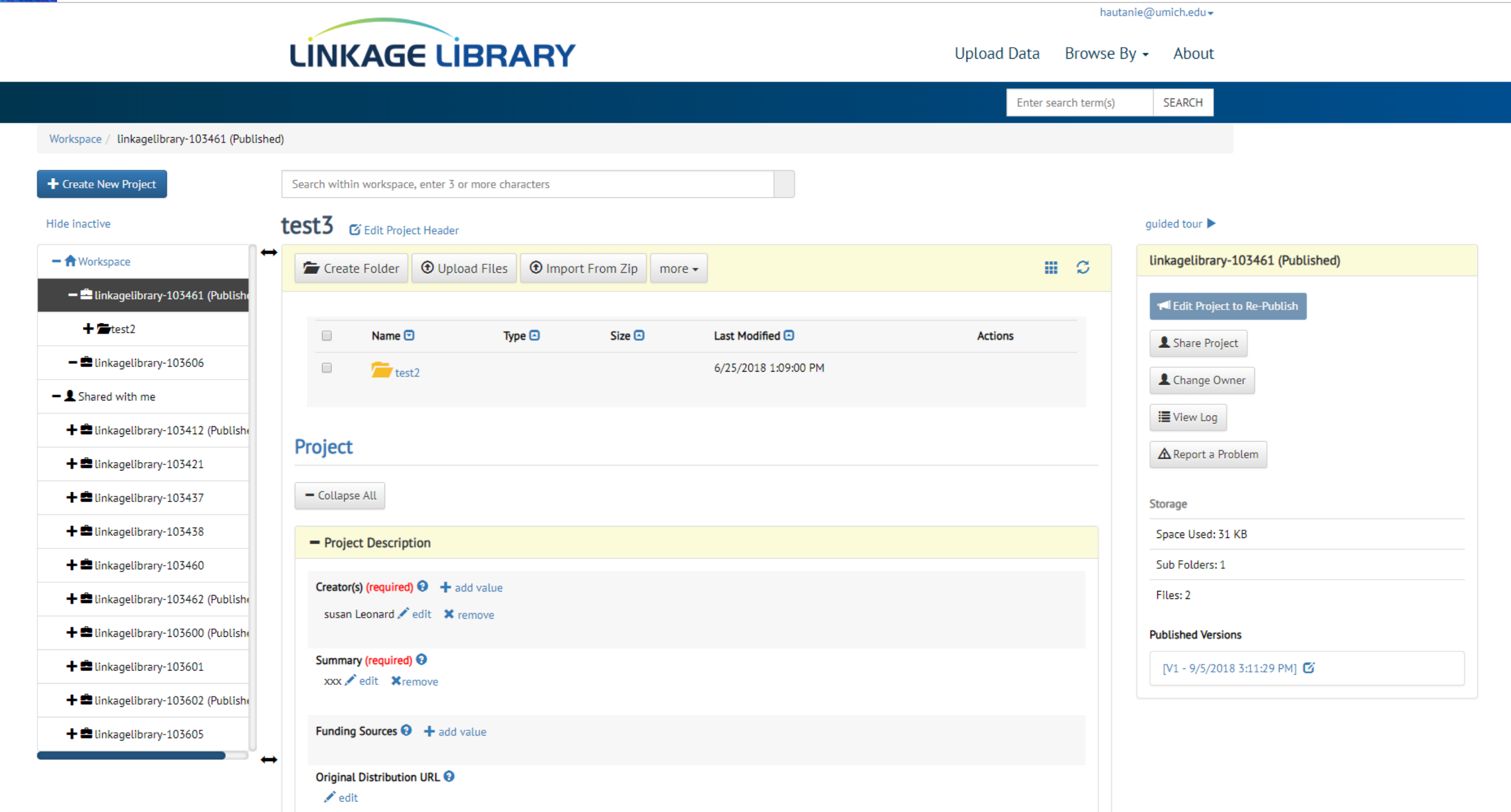
Report a Problem

Storage

Space Used: 2 KB

Sub Folders: 0

Files: 1



Workspace / linkagelibrary-103461 (Published)

- Workspace
  - linkagelibrary-103461 (Publishe
    - test2
  - linkagelibrary-103606
- Shared with me
  - linkagelibrary-103412 (Publishe
  - linkagelibrary-103421
  - linkagelibrary-103437
  - linkagelibrary-103438
  - linkagelibrary-103460
  - linkagelibrary-103462 (Publishe
  - linkagelibrary-103600 (Publishe
  - linkagelibrary-103601
  - linkagelibrary-103602 (Publishe
  - linkagelibrary-103605

test3 Edit Project Header

□	Name	Type	Size	Last Modified	Actions
□	test2			6/25/2018 1:09:00 PM	

Project

Project Description

Creator(s) (required) add value
susan Leonard edit remove

Summary (required)
xxx edit remove

Funding Sources add value

Original Distribution URL
edit

linkagelibrary-103461 (Published)

Storage

Space Used: 31 KB

Sub Folders: 1

Files: 2

Published Versions

[V1 - 9/5/2018 3:11:29 PM]

LINKAGE LIBRARY

Maintaining datasets to support the data linkage community

Email us at linkagelibrary@umich.edu
for queries and updates