ICPSR

SHARING DATA TO ADVANCE SCIENCE

Transparency, reproducibility, and replicability
New initiatives at ICPSR

Margaret C. Levenstein

Director, Inter-university Consortium for Political and Social Research

Promoting transparency and reproducibility

- What is ICPSR?
  - Preserving and accessing shared data and data-related content
  - Journal repositories and journal-related deposits
- Three new initiatives
  - LinkageLibrary
  - SOMAR
  - Researcher passport

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
  - 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

Preserving and accessing shared data and data-related content

- Make data sharing feasible
  - ICPSR's General Archive
    - Anyone can deposit
    - Curated and preserved
  - Guidance over data life cycle
    - Templates for consent, IRB, DMP consistent with transparent and reproducible access
- Incentivize data sharing
    - Standard citation
    - Bibliography
    - Usage statistics

Data in the wild

- Organic or non-designed data creates new challenges for reproducibility and transparency
- Data changes in real time
  - Reproducibility requires snapshots, versioning
- No survey instrument or documentation of sampling to provide metadata to assist in re-use or discovery
  - Or even informed use of data the first time
  - Requires development of standards (extending DDI)
  - Citizen scientists necessary for transparency and reproducibility

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

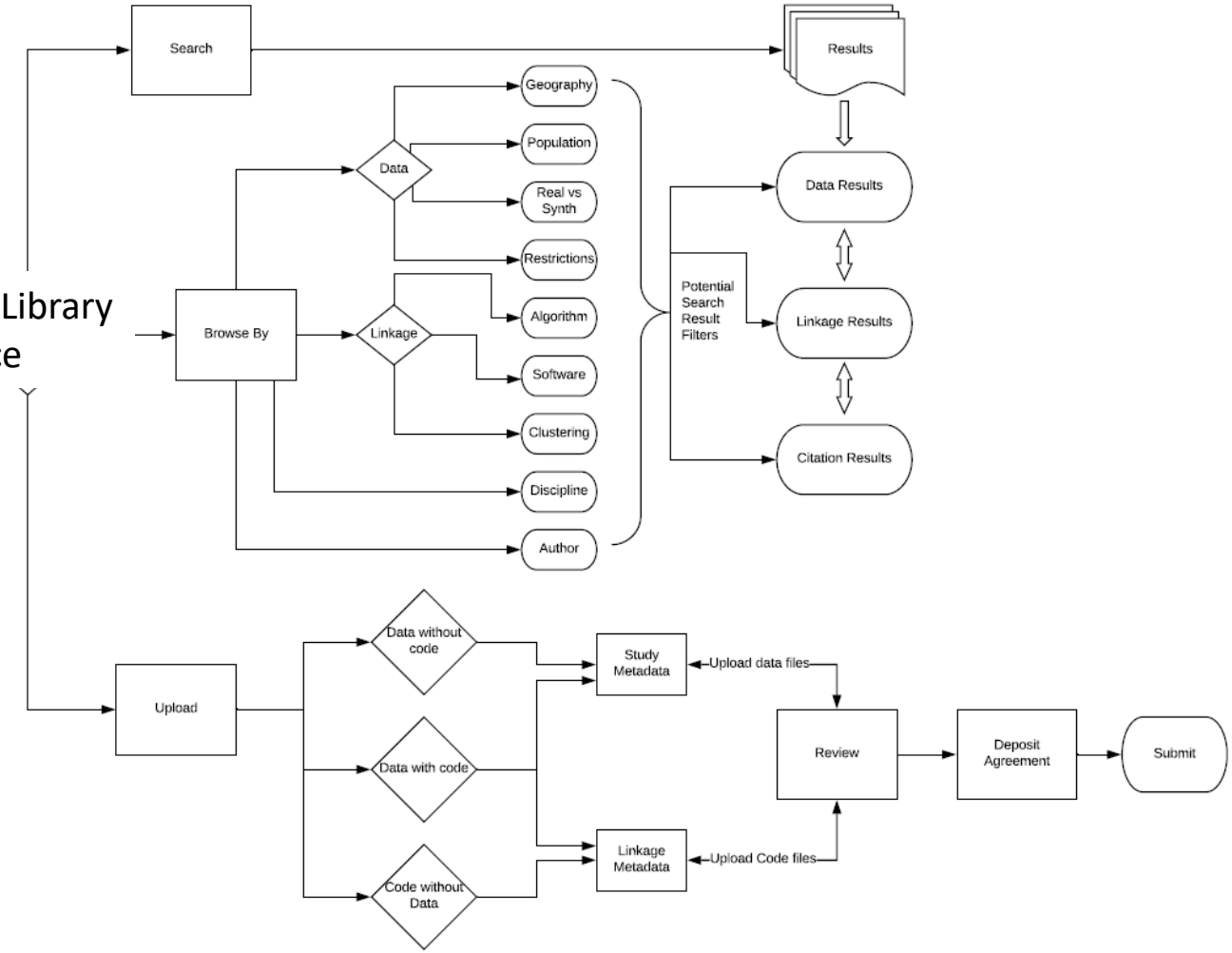
LINKAGE LIBRARY
Maintaining datasets to support the data linkage community

LINKAGE LIBRARY

- Encourage researchers to share linked (or linkable) data, and linkage strategies
  - Algorithms, code
- Compare approaches across projects, datasets, disicplines
  - Improve linkage practices
  - Improve transparency
- Coming this fall
  - Looking for beta testers now!

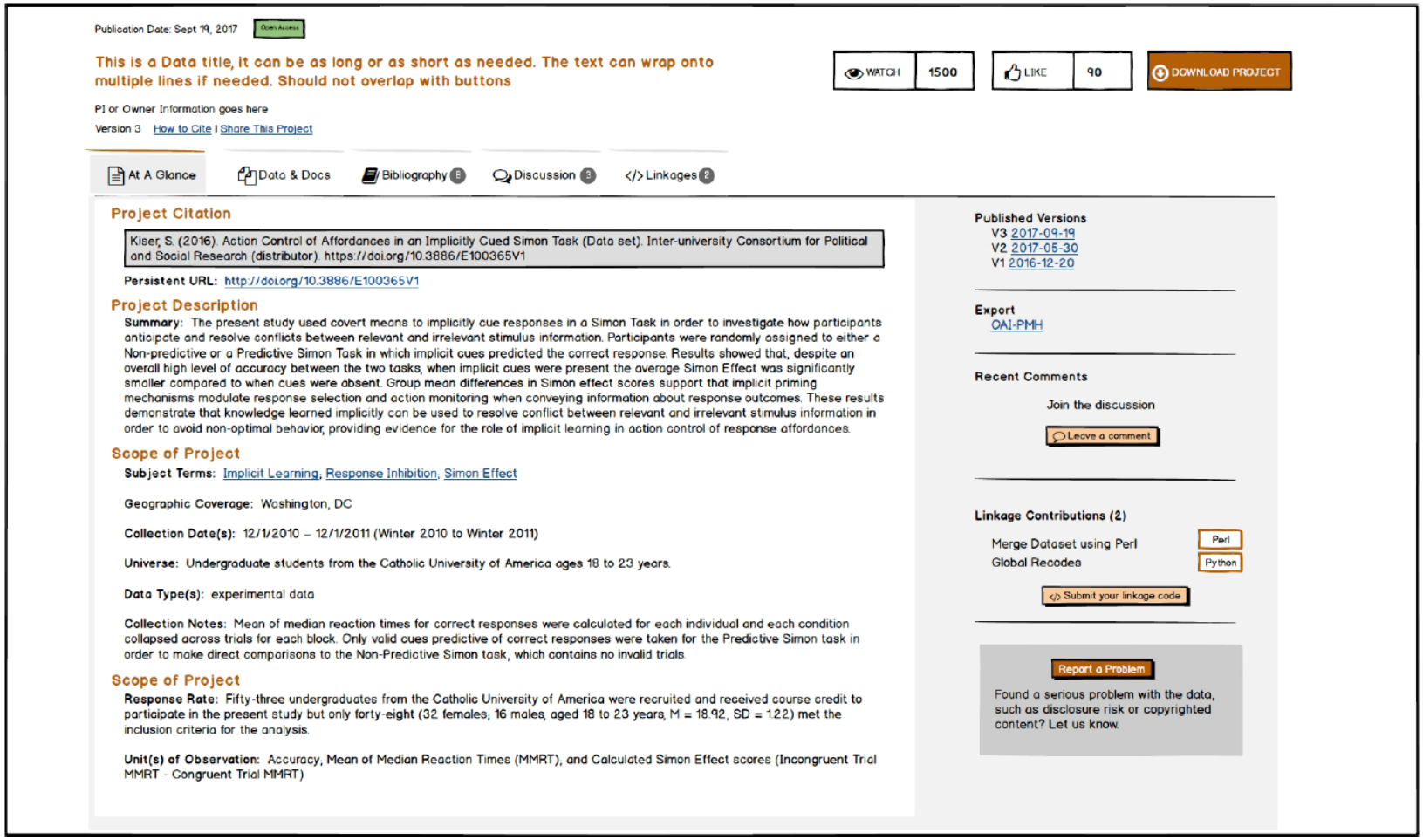
LINKAGE LIBRARY

LinkageLibrary entrance



LINKAGE LIBRARY

Project page



Publication Date: Sept 19, 2017 Open Access

This is a Data title, it can be as long or as short as needed. The text can wrap onto multiple lines if needed. Should not overlap with buttons

WATCH 1500 | LIKE 90 | DOWNLOAD PROJECT

PI or Owner Information goes here

Version 3 How to Cite | Share This Project

At A Glance | Data & Docs | Bibliography 8 | Discussion 3 | Linkages 2

Project Citation

Kiser, S. (2016). Action Control of Affordances in an Implicitly Cued Simon Task (Data set). Inter-university Consortium for Political and Social Research (distributor). https://doi.org/10.3886/E100365V1

Persistent URL: http://doi.org/10.3886/E100365V1

Project Description

Summary: The present study used covert means to implicitly cue responses in a Simon Task in order to investigate how participants anticipate and resolve conflicts between relevant and irrelevant stimulus information. Participants were randomly assigned to either a Non-predictive or a Predictive Simon Task in which implicit cues predicted the correct response. Results showed that, despite an overall high level of accuracy between the two tasks, when implicit cues were present the average Simon Effect was significantly smaller compared to when cues were absent. Group mean differences in Simon effect scores support that implicit priming mechanisms modulate response selection and action monitoring when conveying information about response outcomes. These results demonstrate that knowledge learned implicitly can be used to resolve conflict between relevant and irrelevant stimulus information in order to avoid non-optimal behavior, providing evidence for the role of implicit learning in action control of response affordances.

Scope of Project

Subject Terms: Implicit Learning, Response Inhibition, Simon Effect

Geographic Coverage: Washington, DC

Collection Date(s): 12/1/2010 – 12/1/2011 (Winter 2010 to Winter 2011)

Universe: Undergraduate students from the Catholic University of America ages 18 to 23 years.

Data Type(s): experimental data

Collection Notes: Mean of median reaction times for correct responses were calculated for each individual and each condition collapsed across trials for each block. Only valid cues predictive of correct responses were taken for the Predictive Simon task in order to make direct comparisons to the Non-Predictive Simon task, which contains no invalid trials.

Scope of Project

Response Rate: Fifty-three undergraduates from the Catholic University of America were recruited and received course credit to participate in the present study but only forty-eight (32 females, 16 males, aged 18 to 23 years, M = 18.92, SD = 1.22) met the inclusion criteria for the analysis.

Unit(s) of Observation: Accuracy, Mean of Median Reaction Times (MMRT), and Calculated Simon Effect scores (Incongruent Trial MMRT - Congruent Trial MMRT)

Published Versions
V3 2017-09-19
V2 2017-05-30
V1 2016-12-20

Export
OAI-PMH

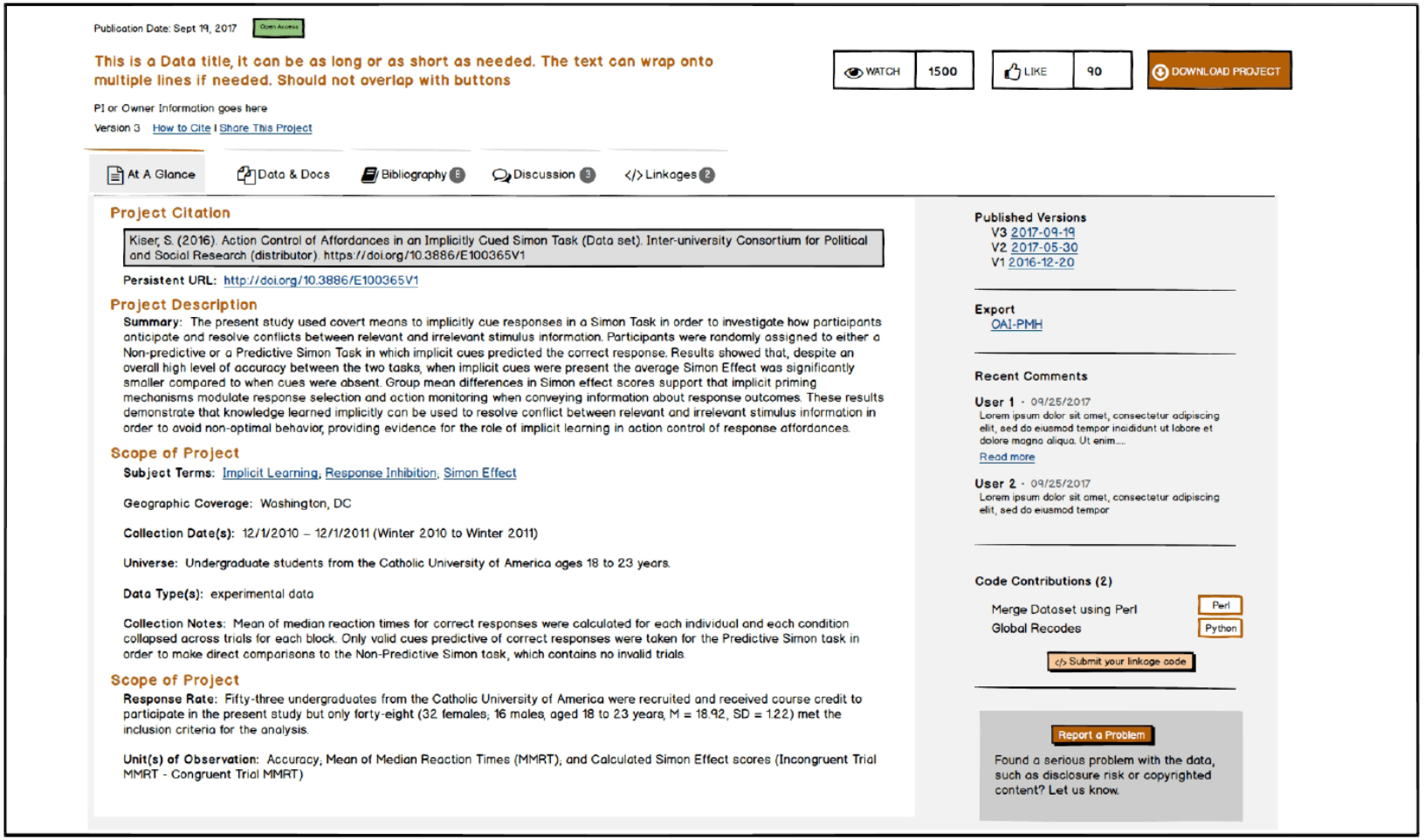
Recent Comments
Join the discussion
Leave a comment

Linkage Contributions (2)
Merge Dataset using Perl — Perl
Global Recodes — Python
Submit your linkage code

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

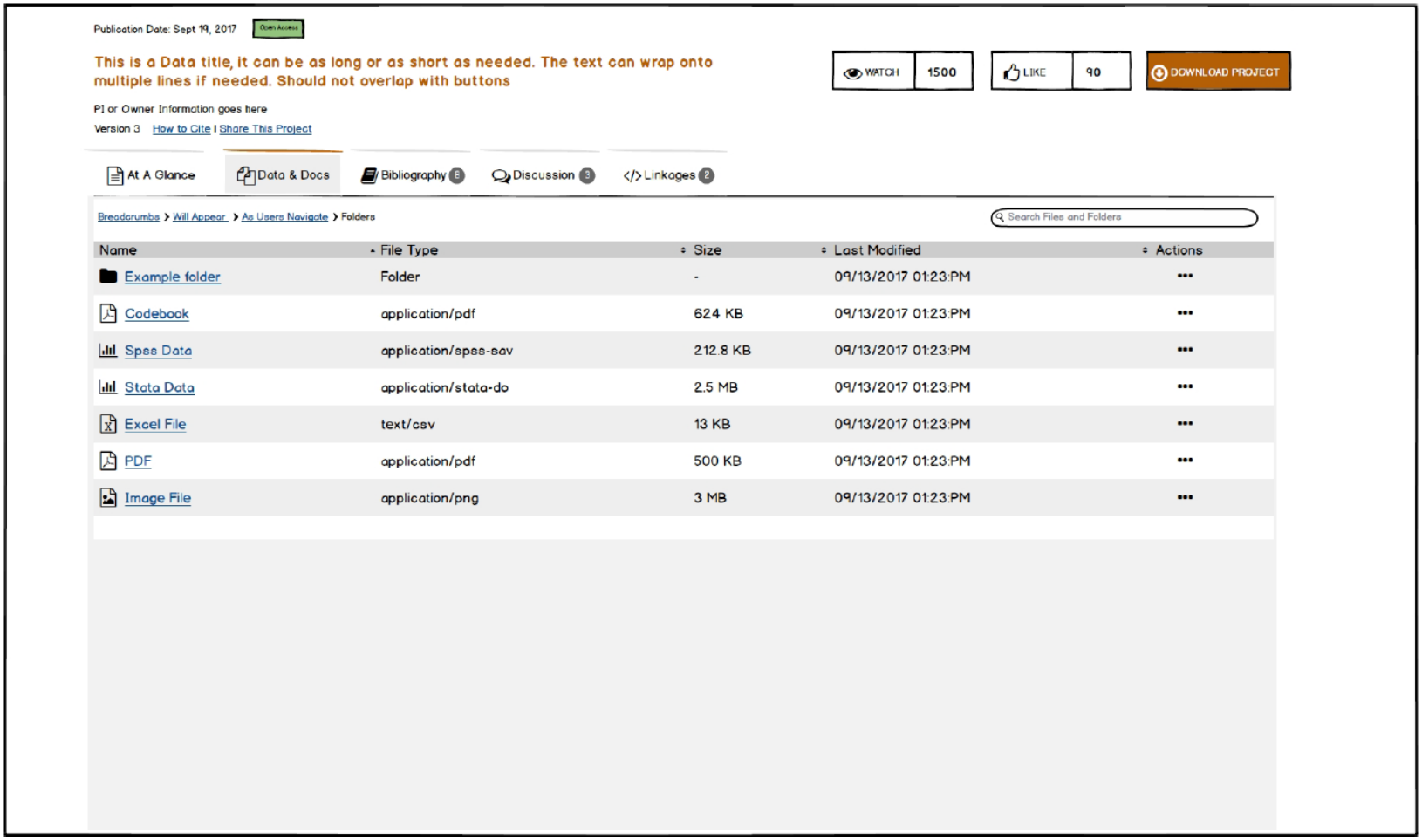
LINKAGE LIBRARY

Others can comment



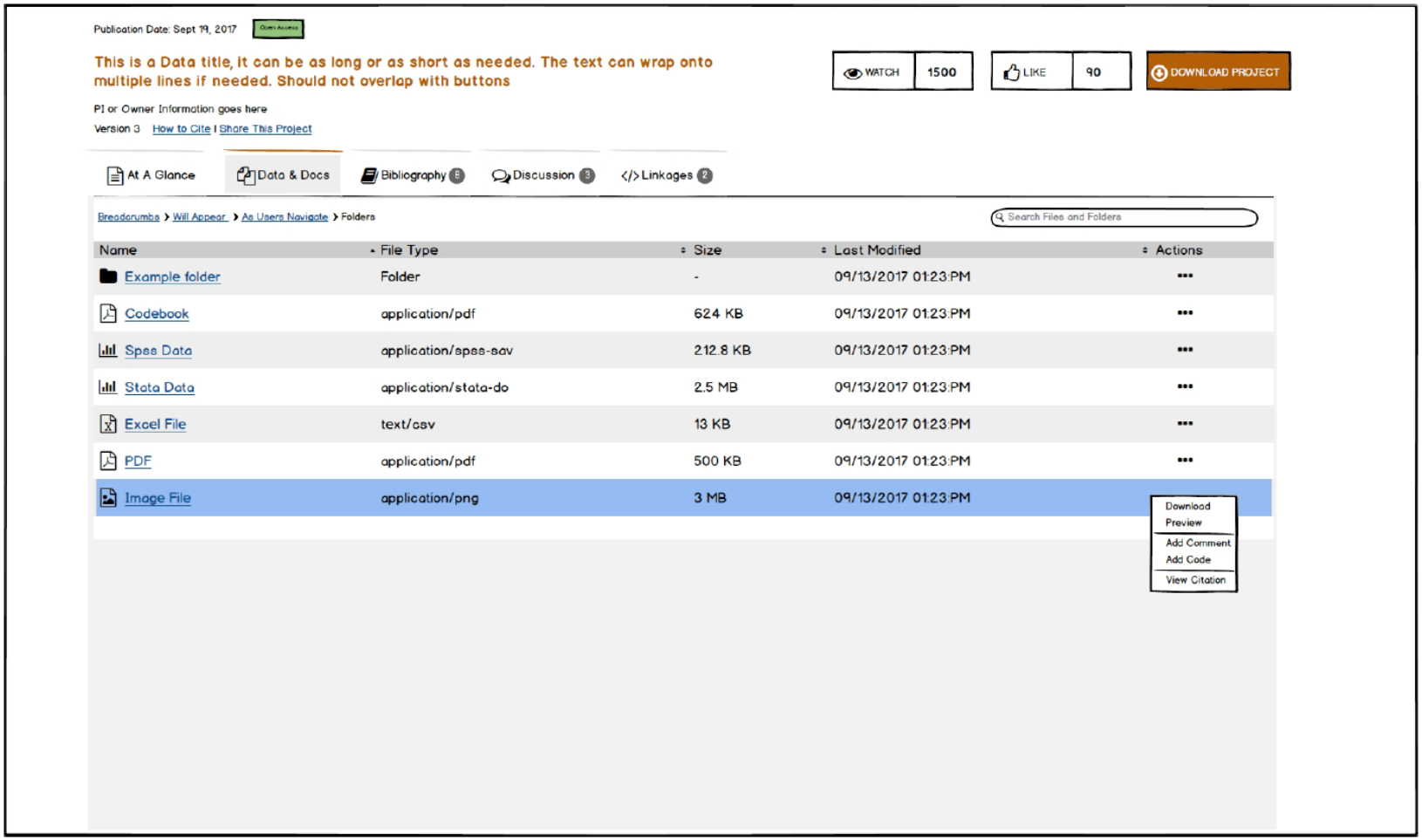
LINKAGE LIBRARY

Share data, documents, code



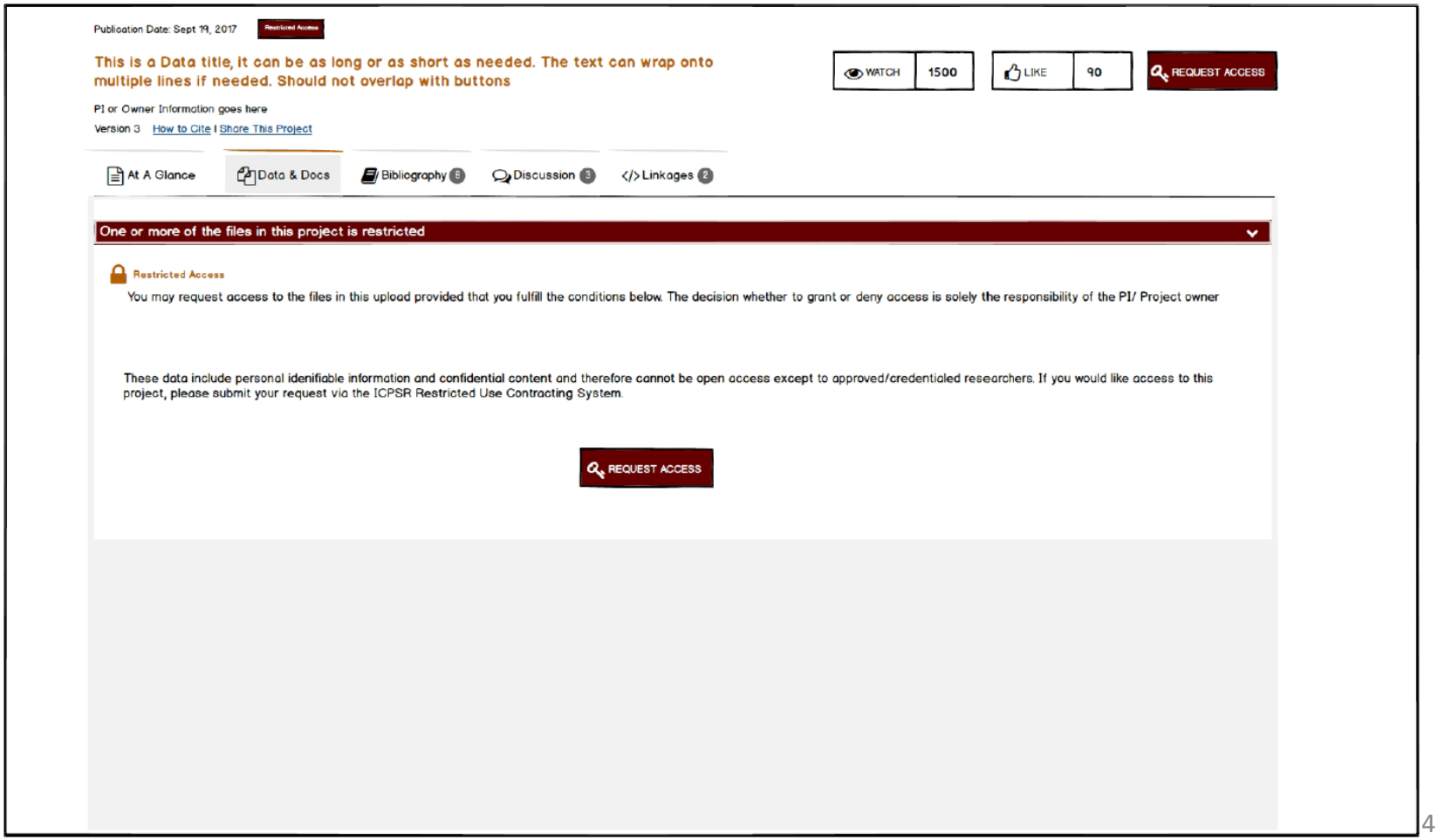
LINKAGE LIBRARY

Share ideas



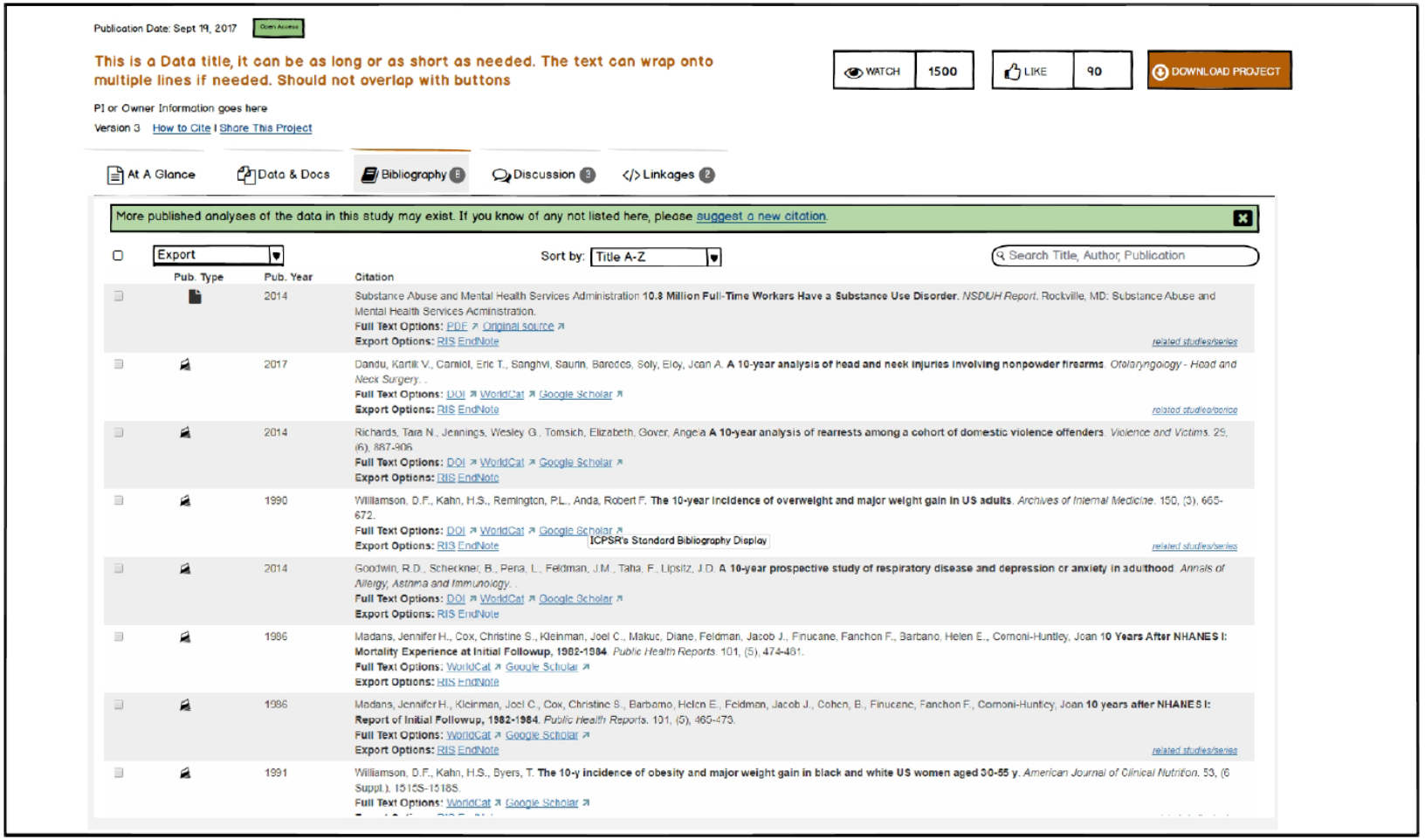
LINKAGE LIBRARY

Protect data and files



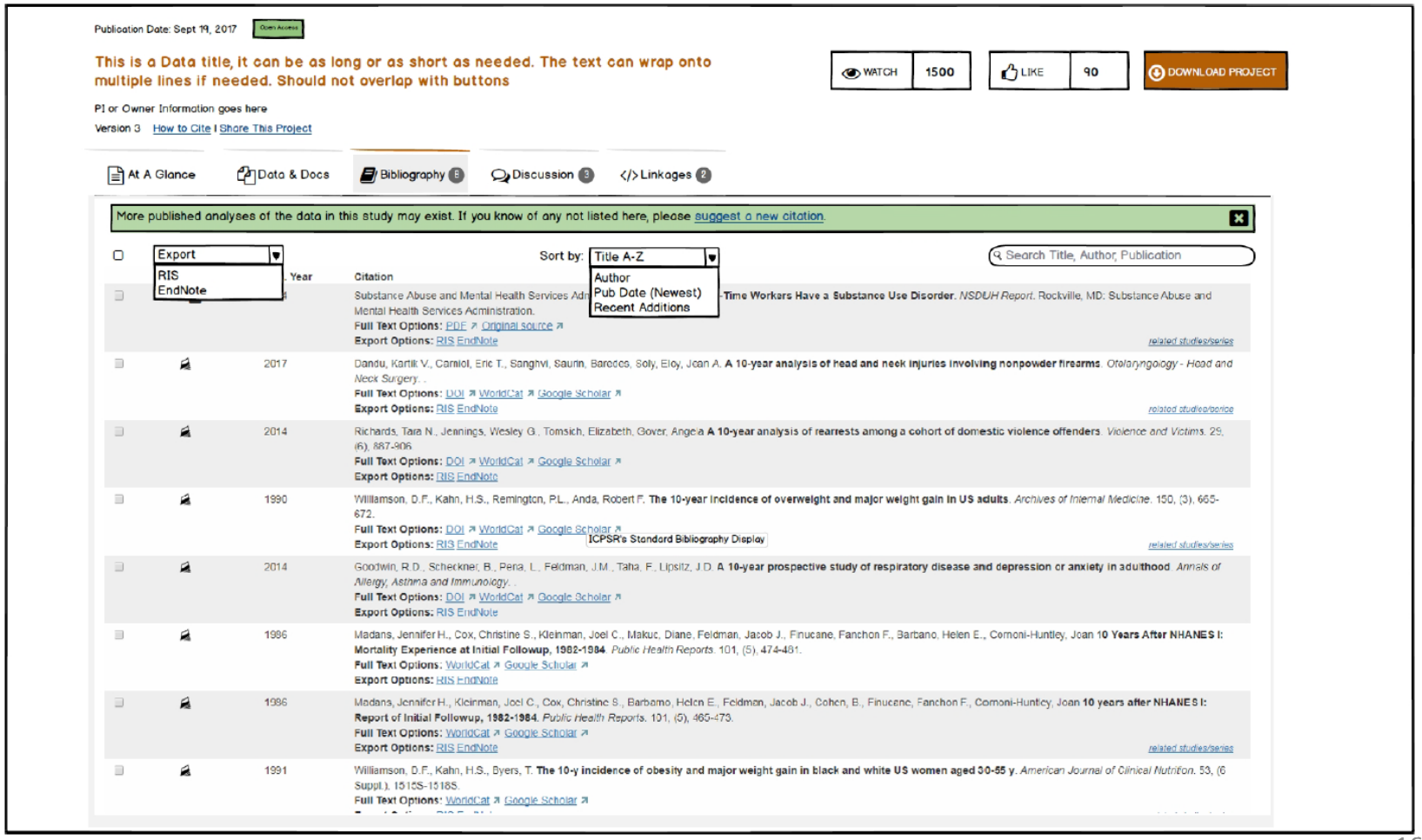
LINKAGE LIBRARY

Bibliography



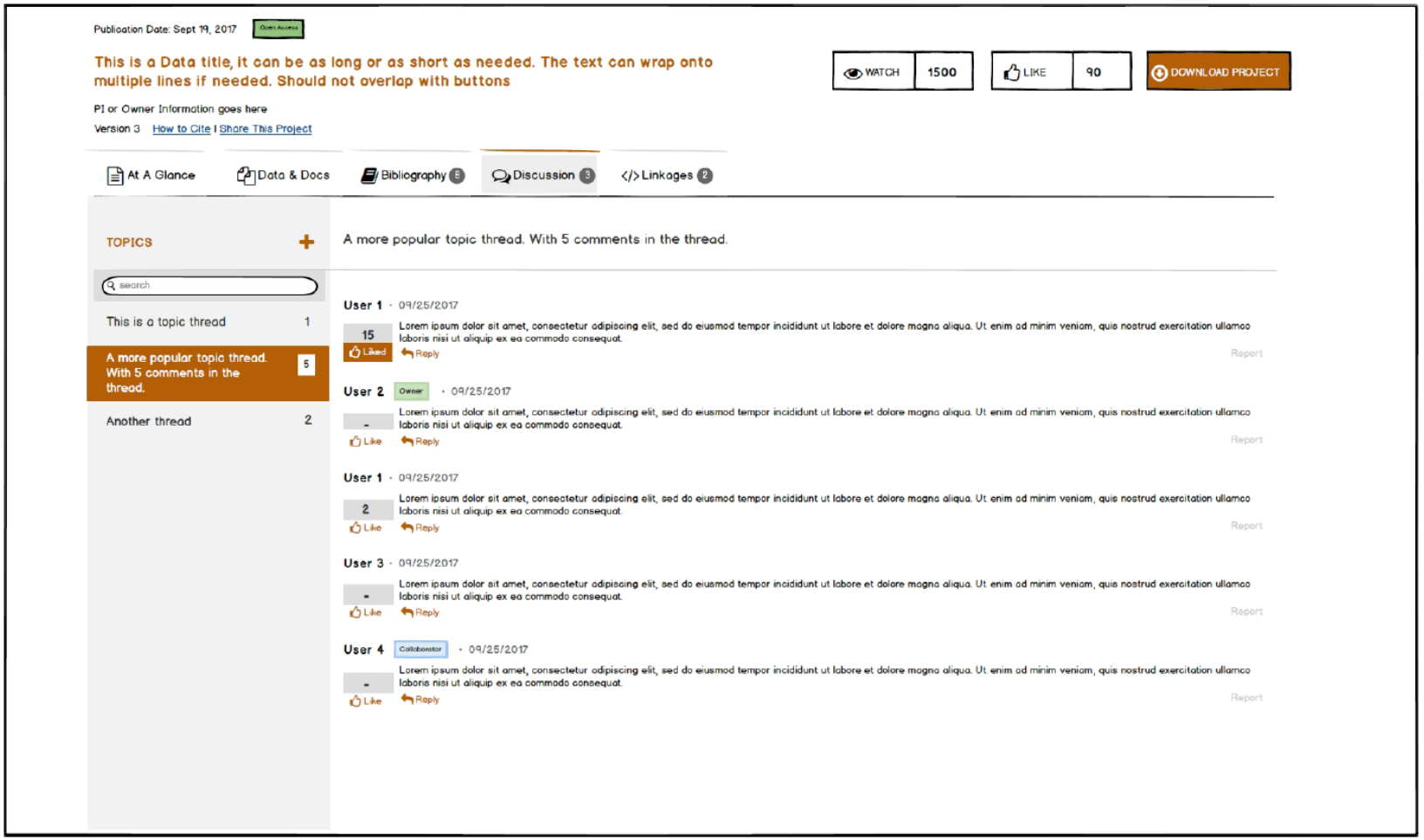
LINKAGE LIBRARY

Bibliography



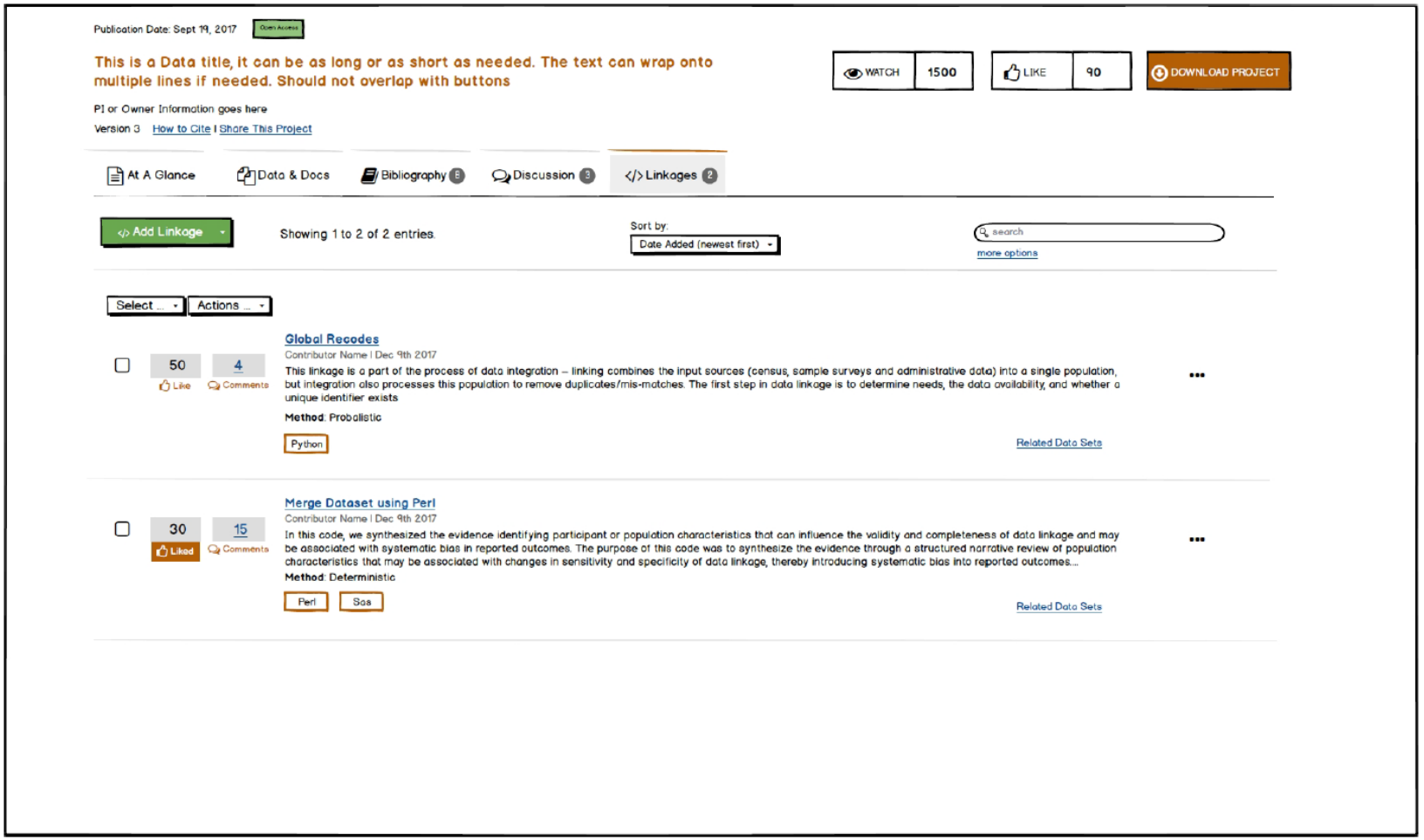
LINKAGE LIBRARY

Threaded discussion



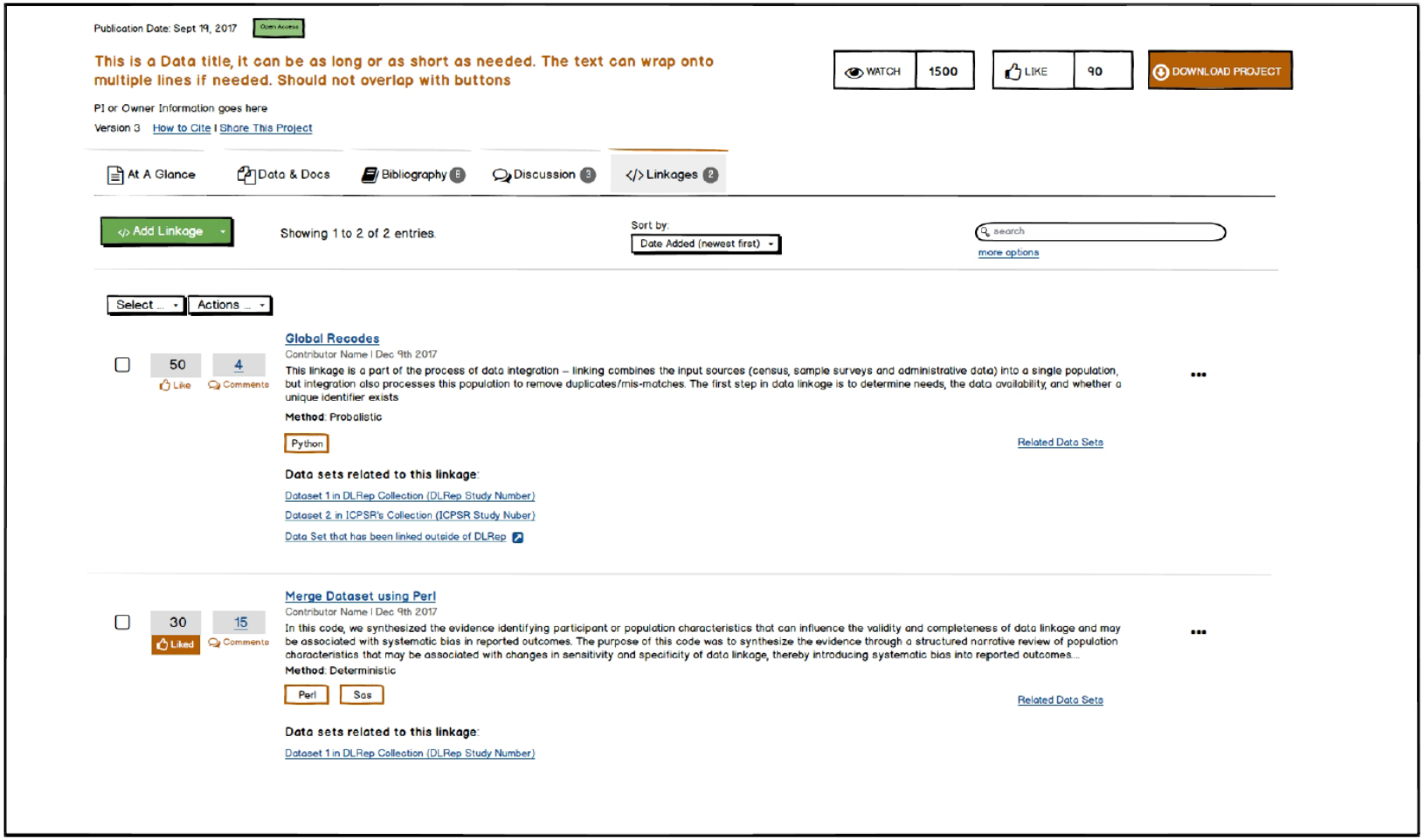
LINKAGE LIBRARY

Linkage algorithms



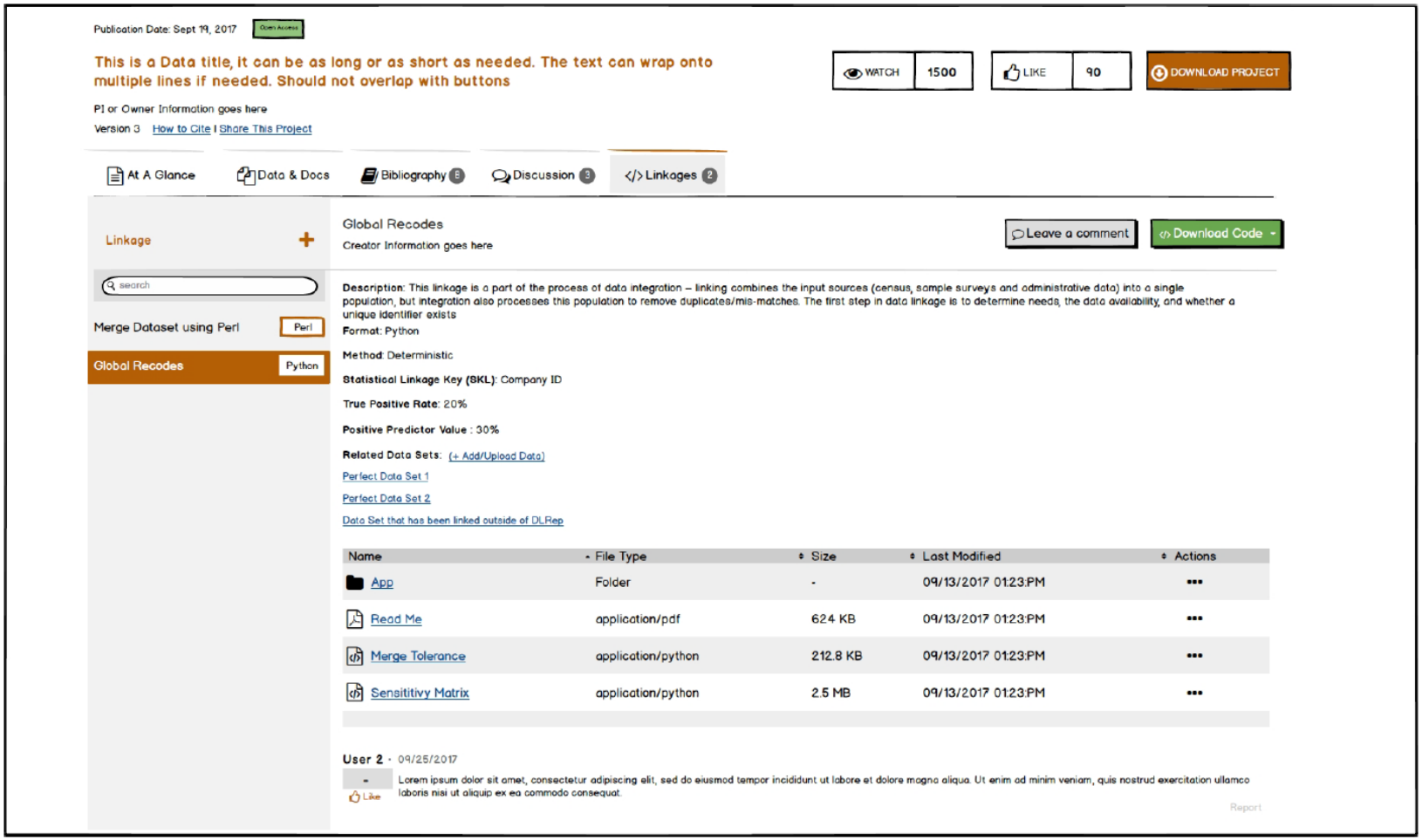
LINKAGE LIBRARY

Linkage algorithms



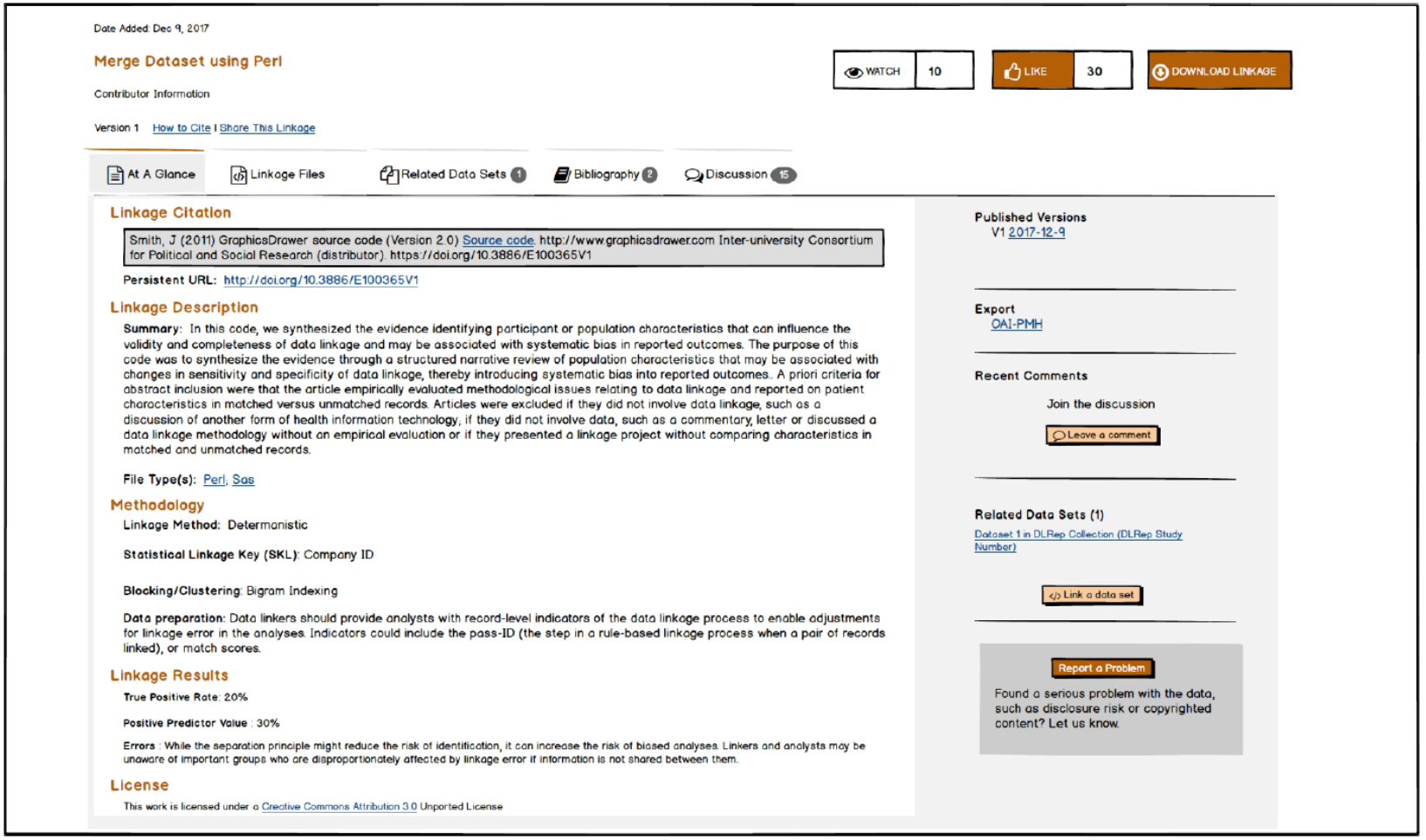
LINKAGE LIBRARY

Related data & documents



LINKAGE LIBRARY

More about linkage strategy



Date Added: Dec 9, 2017

Merge Dataset using Perl

WATCH 10 | LIKE 30 | DOWNLOAD LINKAGE

Contributor Information

Version 1 How to Cite | Share This Linkage

At A Glance | Linkage Files | Related Data Sets 1 | Bibliography 2 | Discussion 15

Linkage Citation

Smith, J (2011) GraphicsDrawer source code (Version 2.0) Source code. http://www.graphicsdrawer.com Inter-university Consortium for Political and Social Research (distributor). https://doi.org/10.3886/E100365V1

Persistent URL: http://doi.org/10.3886/E100365V1

Linkage Description

Summary: In this code, we synthesized the evidence identifying participant or population characteristics that can influence the validity and completeness of data linkage and may be associated with systematic bias in reported outcomes. The purpose of this code was to synthesize the evidence through a structured narrative review of population characteristics that may be associated with changes in sensitivity and specificity of data linkage, thereby introducing systematic bias into reported outcomes. A priori criteria for abstract inclusion were that the article empirically evaluated methodological issues relating to data linkage and reported on patient characteristics in matched versus unmatched records. Articles were excluded if they did not involve data linkage, such as a discussion of another form of health information technology, if they did not involve data, such as a commentary, letter or discussed a data linkage methodology without an empirical evaluation or if they presented a linkage project without comparing characteristics in matched and unmatched records.

File Type(s): Perl, Sas

Methodology

Linkage Method: Deterministic

Statistical Linkage Key (SKL): Company ID

Blocking/Clustering: Bigram Indexing

Data preparation: Data linkers should provide analysts with record-level indicators of the data linkage process to enable adjustments for linkage error in the analyses. Indicators could include the pass-ID (the step in a rule-based linkage process when a pair of records linked), or match scores.

Linkage Results

True Positive Rate: 20%

Positive Predictor Value : 30%

Errors : While the separation principle might reduce the risk of identification, it can increase the risk of biased analyses. Linkers and analysts may be unaware of important groups who are disproportionately affected by linkage error if information is not shared between them.

License

This work is licensed under a Creative Commons Attribution 3.0 Unported License

Published Versions
V1 2017-12-9

Export
OAI-PMH

Recent Comments

Join the discussion

Leave a comment

Related Data Sets (1)

Dataset 1 in DLRep Collection (DLRep Study Number)

Link a data set

Report a Problem

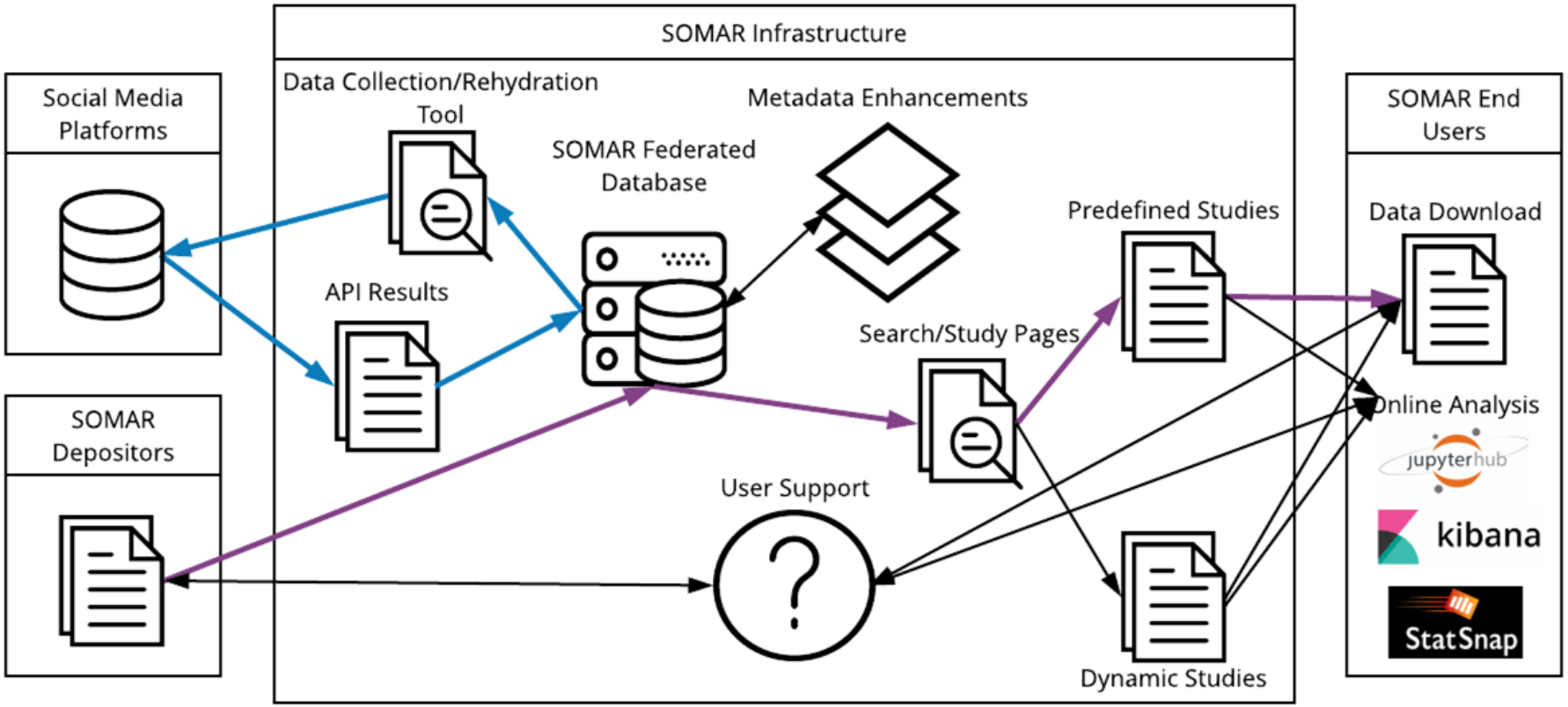
Found a serious problem with the data, such as disclosure risk or copyrighted content? Let us know.

SOMAR: Social Media Archive

- Addresses 4 communities who:
  - study social media use specifically
  - leverage social media data to understand people and society
  - study social science methods
  - Investigate new methods for curation, publication, confidentiality and quality assessment, and long-term management of research data
- Archive enables historical and longitudinal analyses often missing from rapidly changing social medial platforms

SOMAR: Social Media Archive

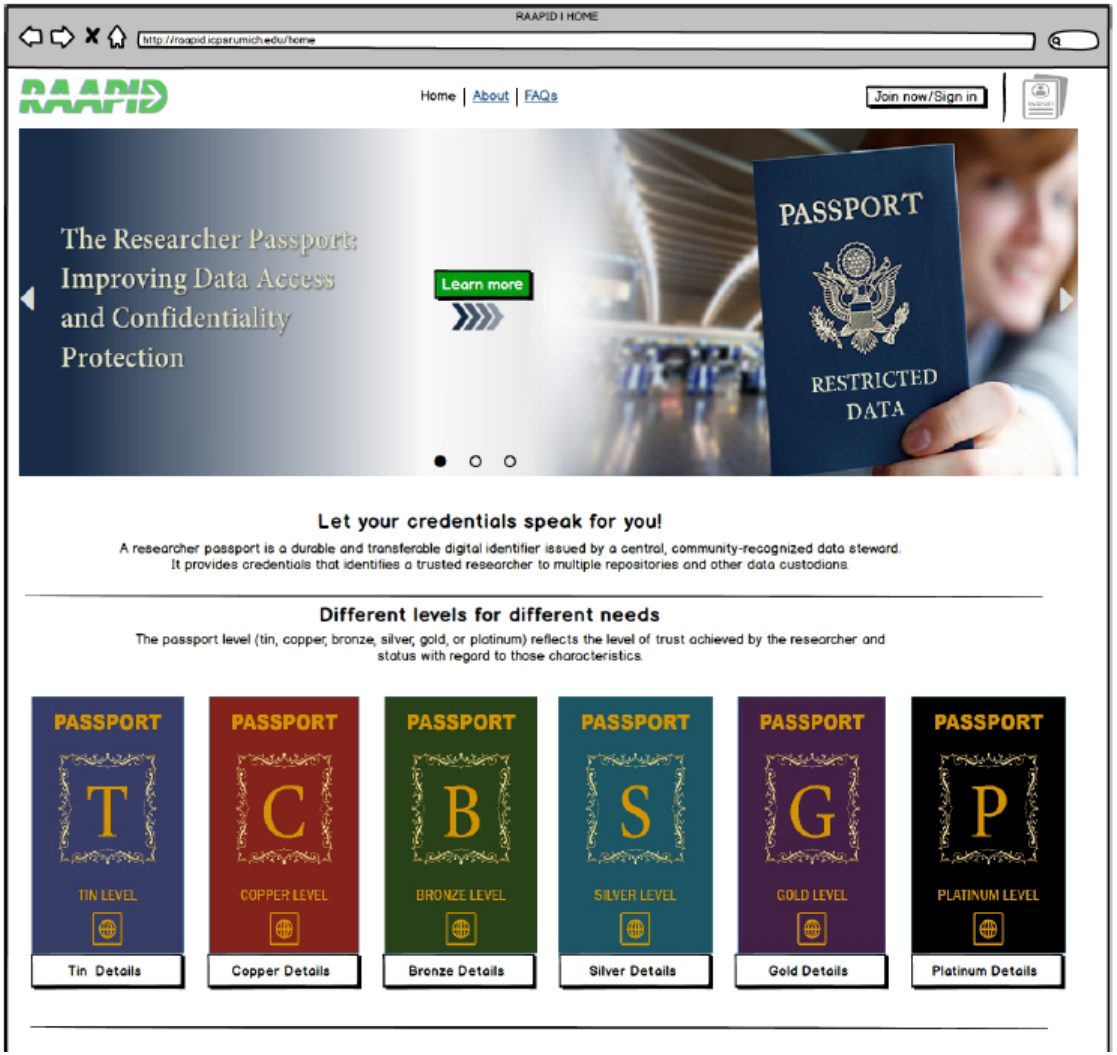
- Archive data where possible
- Archive workflows and code where data sharing is prohibited
  - Eg: Twitter IDs and code for rehydrating
- Curation and metadata
  - Provenance, dates, hashtags, confidentiality protection


SOMAR Infrastructure
Social Media Platforms
Data Collection/Rehydration Tool
Metadata Enhancements
SOMAR Federated Database
API Results
Predefined Studies
Search/Study Pages
SOMAR End Users
Data Download
Online Analysis
jupyterhub
kibana
StatSnap
SOMAR Depositors
User Support
Dynamic Studies

SOMAR Challenges

- Technical infrastructure
- Ethical and legal infrastructure
- Metadata enhancements
- Adoption

Researcher Passport



Establishing shared understanding of what it means to be a trusted researcher

Researcher Passport

- Researcher Passport: Improving Data Access and Confidentiality Protection
  - ICPSR's Strategy for a Community-normed System of Digital Identities of Access
  - https://deepblue.lib.umich.edu/handle/2027.42/143808
  - Identifies inconsistent language and policies that impede access
  - Facilitate sharing of *proprietary* data
- Passports for safe people
  - Verified identities, institutional affiliation
  - Training
  - Experience (good and bad)
- Visas to control access
  - Permission to "enter" (access) specific data specifying
    - Passport holder
      - Project, Place, Period

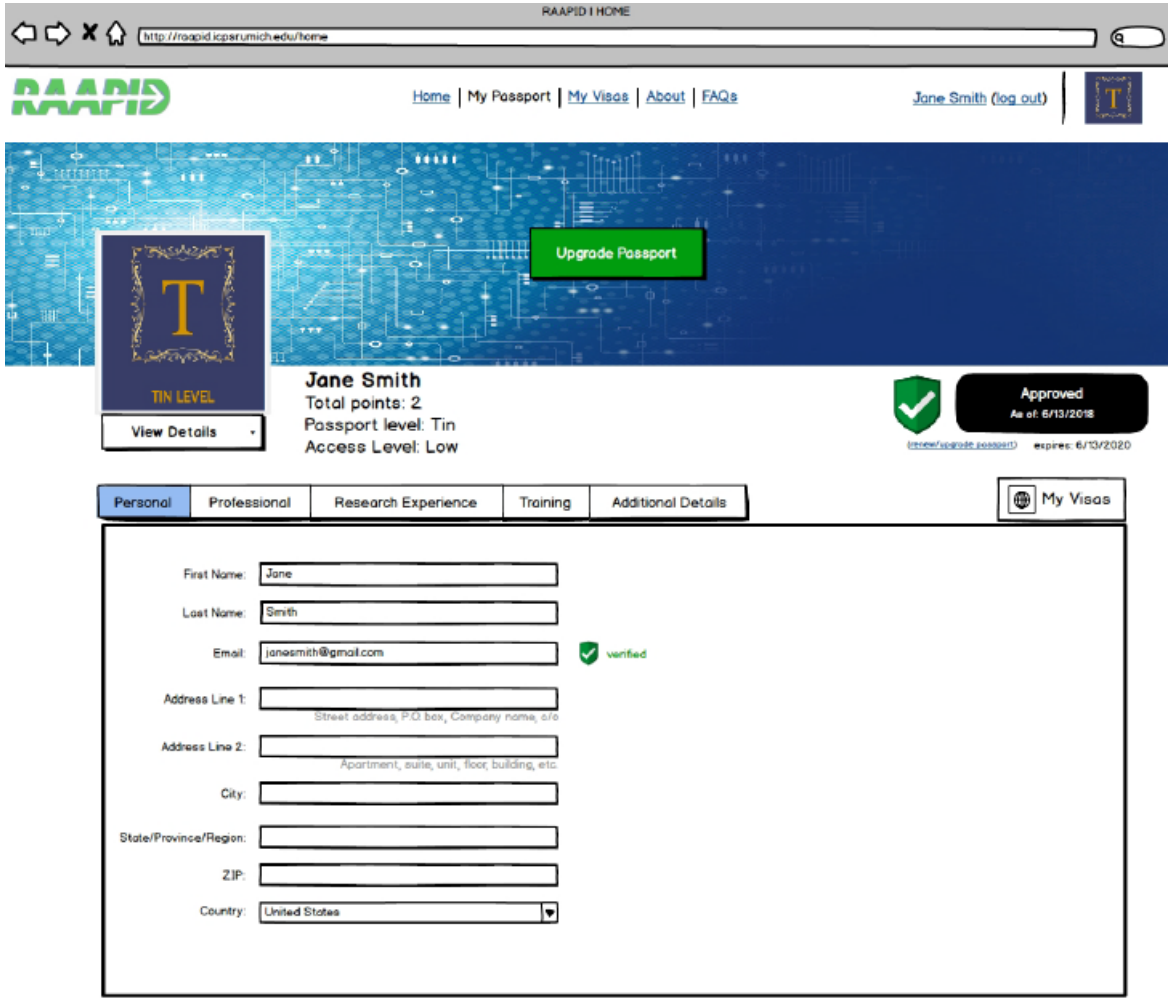
Passport levels

PI SCORE	LOW	MODERATE	HIGH	HIGHEST	
	DATA SECURITY LEVEL				
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin

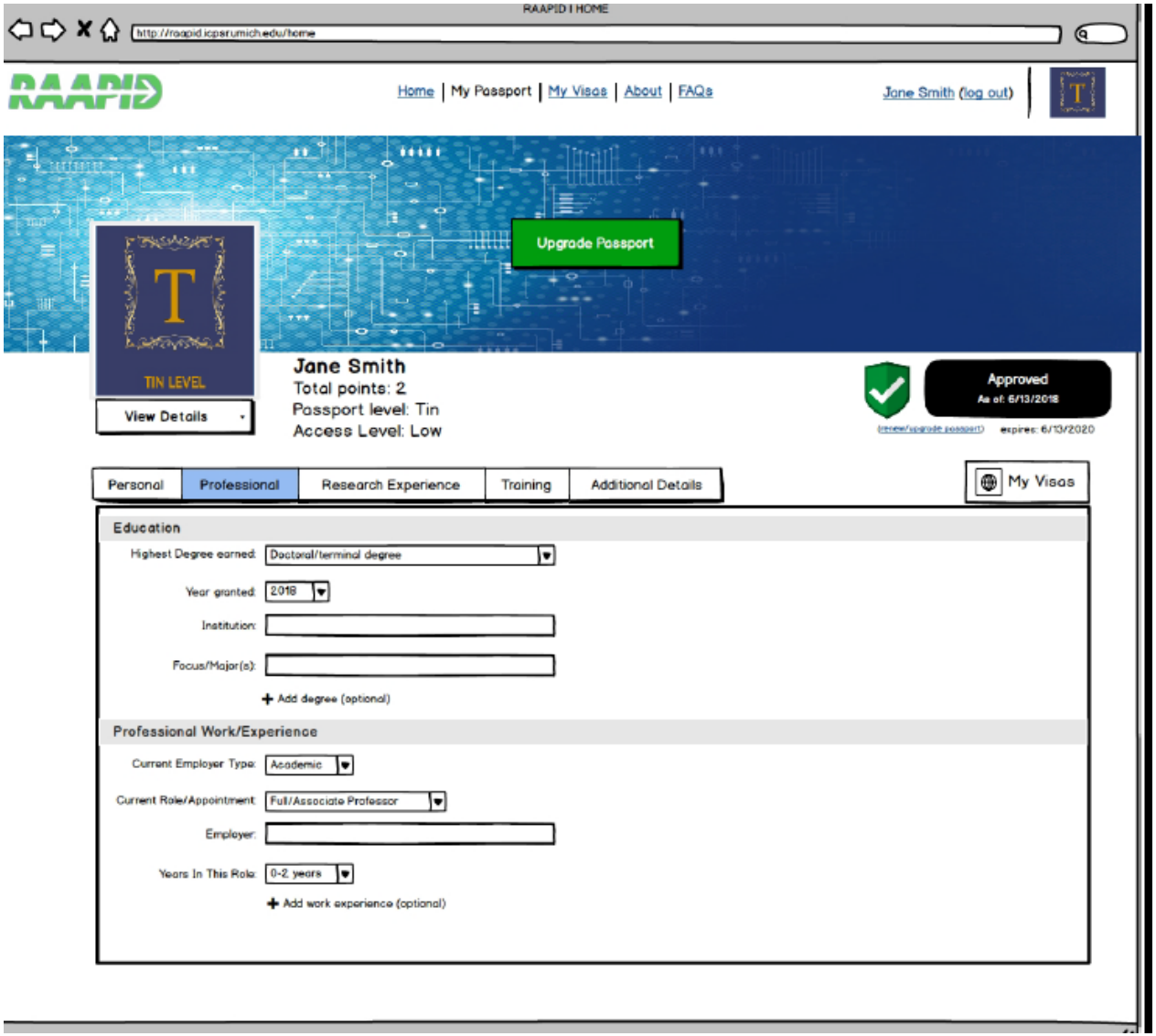
Researcher passport

- Implementation plan
  - Issue passports as digital identifiers associated with ICPSR MyData Accounts
    - Other passport issuers?
    - Blockchain
  - Repositories issue visas to access restricted data
    - Visa controls access to secure download or VDE
      - Visa issued after signed DUA
  - Establishing content of training requirements
    - Determining which existing training include that content
  - Governance system to provide due process for allegations of misconduct

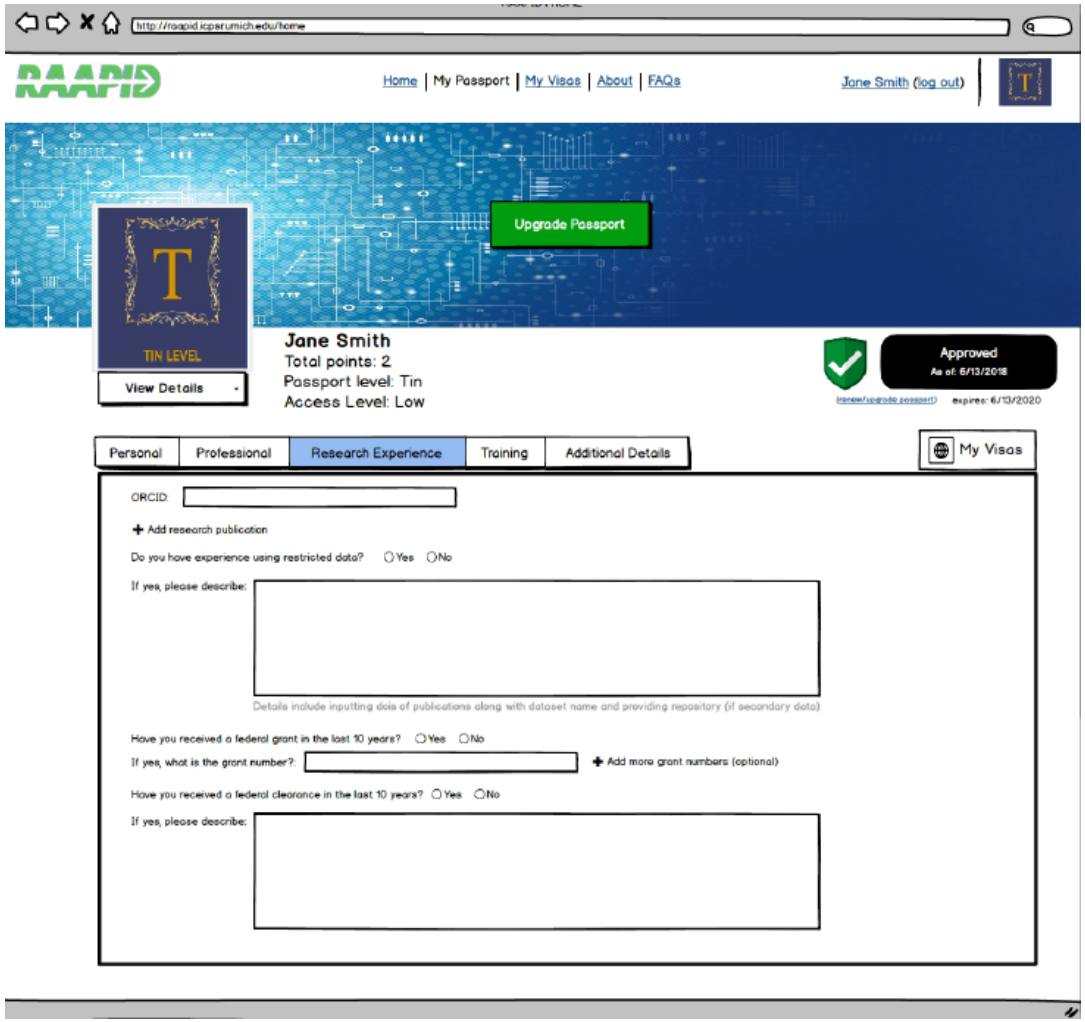
On-line passport application
Verified identification

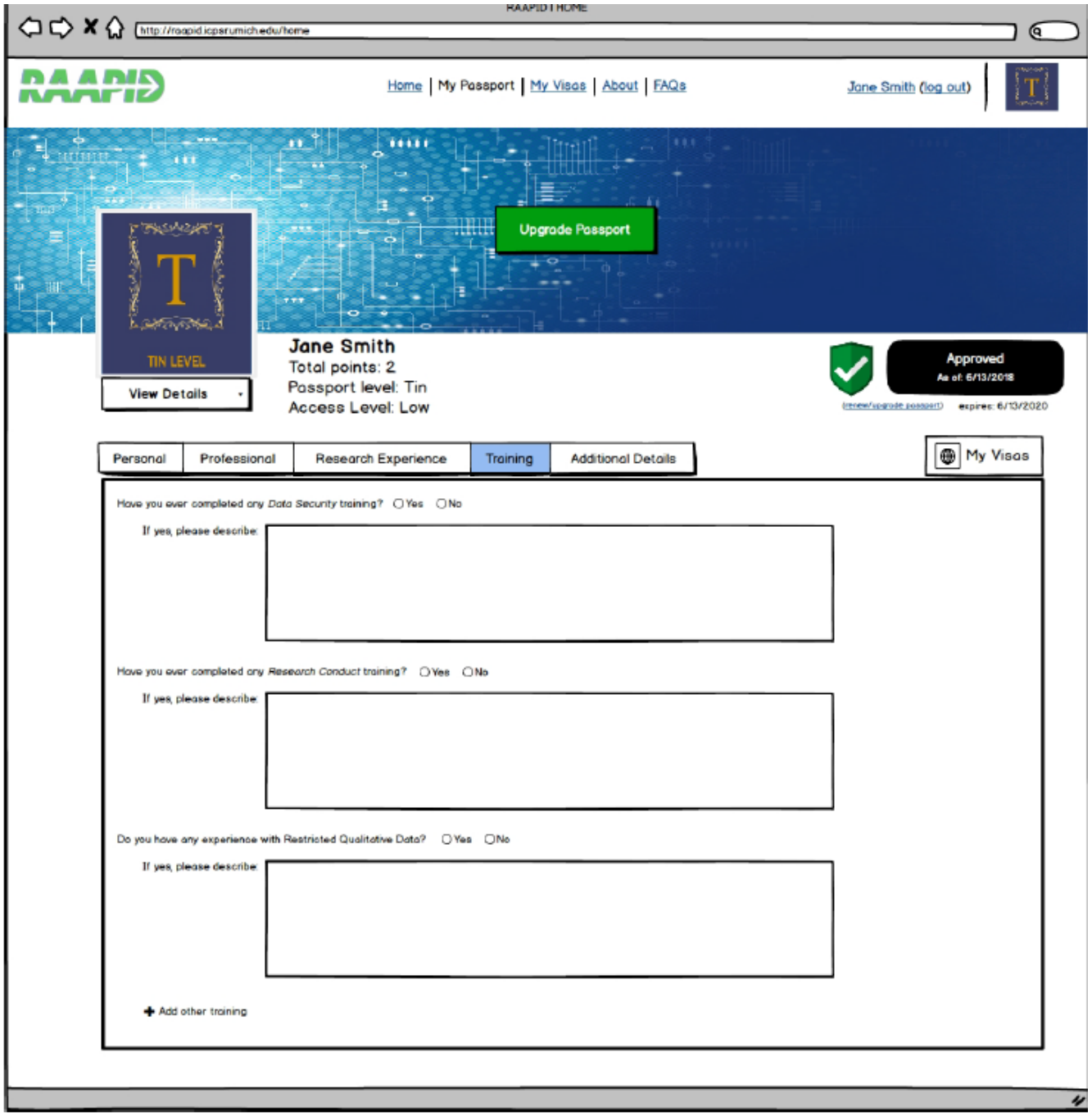


Professional qualifications and affiliations

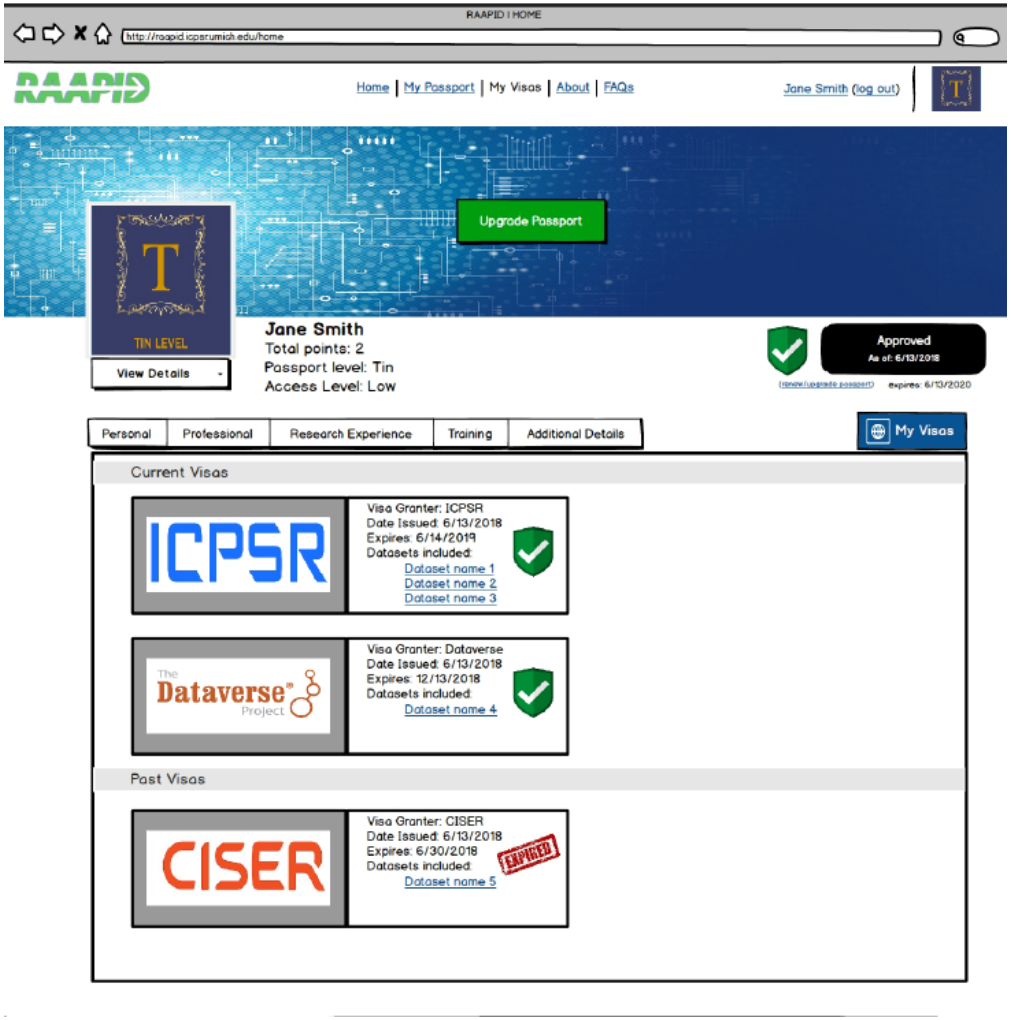


History of data stewardship




RAAPID | HOME
http://raapid.icpsr.umich.edu/home
RAAPID
Home | My Passport | My Visas | About | FAQs
Jane Smith (log out)
Upgrade Passport
TIN LEVEL
View Details
Jane Smith
Total points: 2
Passport level: Tin
Access Level: Low
Approved
As of: 6/13/2018
(renew/upgrade passport)
expires: 6/13/2020
Personal
Professional
Research Experience
Training
Additional Details
My Visas
Have you ever completed any Data Security training? Yes No
If yes, please describe:
Have you ever completed any Research Conduct training? Yes No
If yes, please describe:
Do you have any experience with Restricted Qualitative Data? Yes No
If yes, please describe:
Add other training

Participating repository issue visas



Note that this is conceptual. These repositories have not agreed to this!

Questions

- Can reach ICPSR at help@icpsr.umich.edu
- Researcher Credentialing
  - Johanna Bleckman at Bleckman@umich.edu
- LinkageLibrary
  - Susan Leonard at hautanie@umich.edu
- SOMAR
  - Libby Hemphill at LibbyH@umich.edu