Linking databases to code repositories with Throughput

March 3, 20201
WestGrid Virtual Presentation

Simon Goring
Kerstin Lehnert, Nick McKay, Steve Kuehn, Shane Loeffler, Anders Noren, Andrea Thomer, Socorro Dominguez

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Introduction

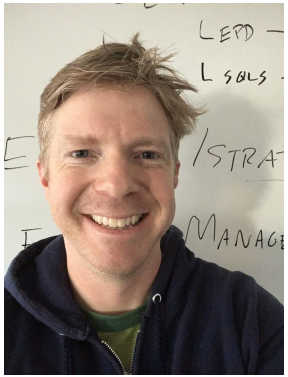
Simon Goring

- Associate Scientist, Geography (UWisc)
- Adjunct Professor, Computer Science (UBC)

Motivated by a desire to help mobilize data to improve equity and reduce “time to science”.

Throughput (https://throughputdb.com)

A three year NSF Funded project through the EarthCube (http://earthcube.org) program to connect data resources in the geosciences and beyond.



Twitter: @sjGoring

Collaborators

Kerstin Lehnert - Columbia University EarthChem/SESAR/IGSNs
Nick McKay - Northern Arizona University (LinkedEarth)
Steve Kuehn - Concord University
Shane Loeffler - University of Minnesota (https://flyovercountry.io/)
Anders Noren - University of Minnesota (CSDCO, LacCore)
Andrea Thomer - University of Michigan School of Information
Socorro Dominguez - University of Wisconsin (UBC Data Science)

Interdisciplinary research is complicated.

Finding Appropriate Data -> Using Appropriate Methods -> Ensuring Reproducibility -> Obtaining Credit

Research is complicated.

Finding Appropriate Data -> Using Appropriate Methods -> Ensuring Reproducibility -> Obtaining Credit

Problems to Solve

1. Finding Appropriate Data
2. Using Appropriate Methods
3. Ensuring Reproducibility
4. Obtaining Credit

Problems to Solve

1. **Finding Appropriate Data**

   Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2. **Using Appropriate Methods**

3. **Ensuring Reproducibility**

4. **Obtaining Credit**

Problems to Solve

1. **Finding Appropriate Data**

   Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2. **Using Appropriate Methods**

   Time series analysis, data processing, R packages.

3. **Ensuring Reproducibility**

4. **Obtaining Credit**

Problems to Solve

1

Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.

3

Ensuring Reproducibility

Finding support, reproducing best practices, sustainability.

4

Obtaining Credit

Problems to Solve

1

Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.

3

Ensuring Reproducibility

Finding support, reproducing best practices, sustainability.

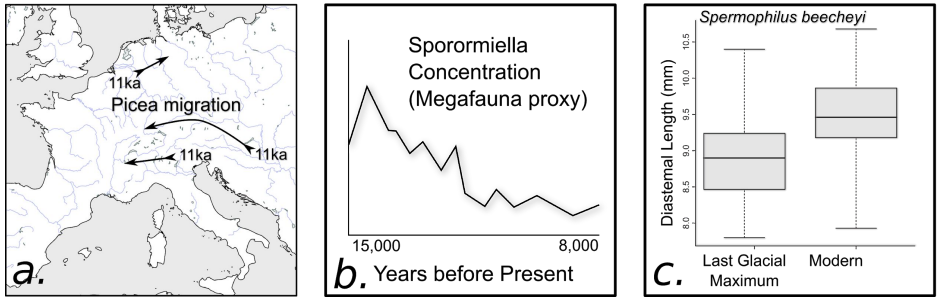
4

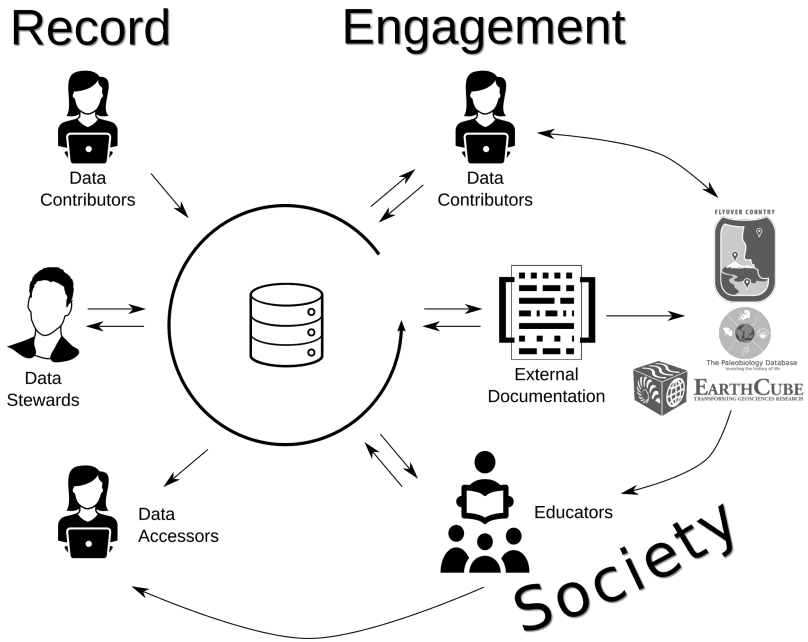
Obtaining Credit

Credit for methods and products beyond papers.

10

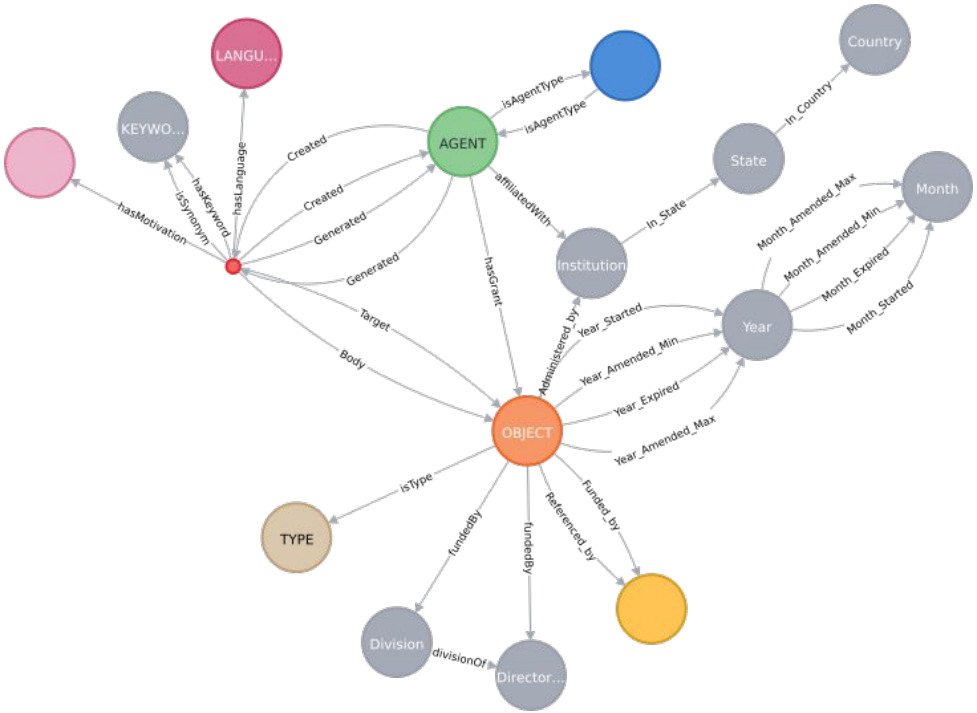
Mid-Scale Infrastructure





Goring SJ, Graham R, Loeffler S, Myrbo A, Oliver JS, Ormond C, & Williams JW. 2018. *The Neotoma Paleoecology Database: A Research Outreach Nexus*. Elements of Paleontology. Cambridge: Cambridge University Press. DOI: **10.1017/9781108681582**

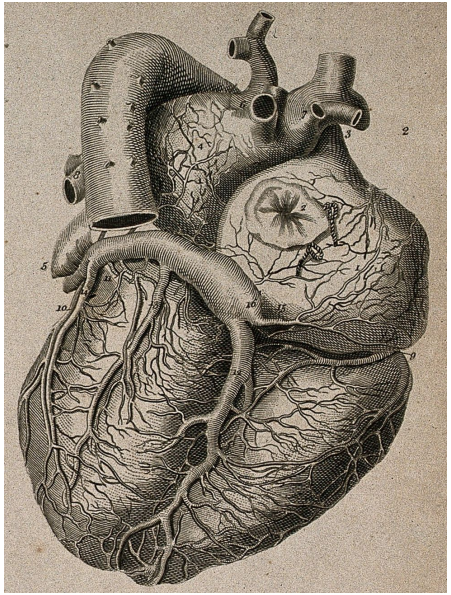
Solutions from a Graph



Throughput provides:

- A platform for data object annotation.
- A tool to discover connections
- A system for managing non-authoratative metadata
- A tool to understand and track data and software citation

At its Heart, Annotations

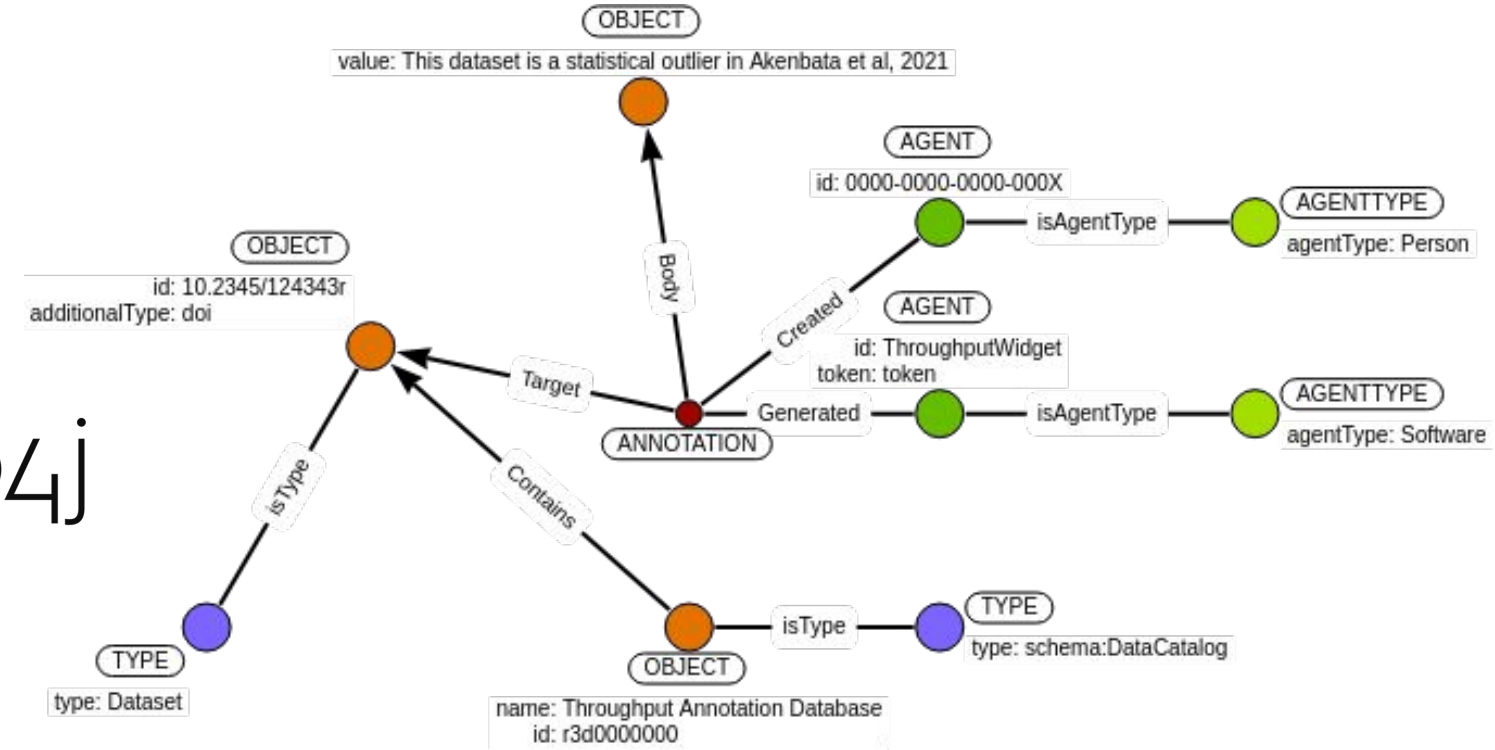


W3C Annotation model (https://www.w3.org/TR/annotation-model)

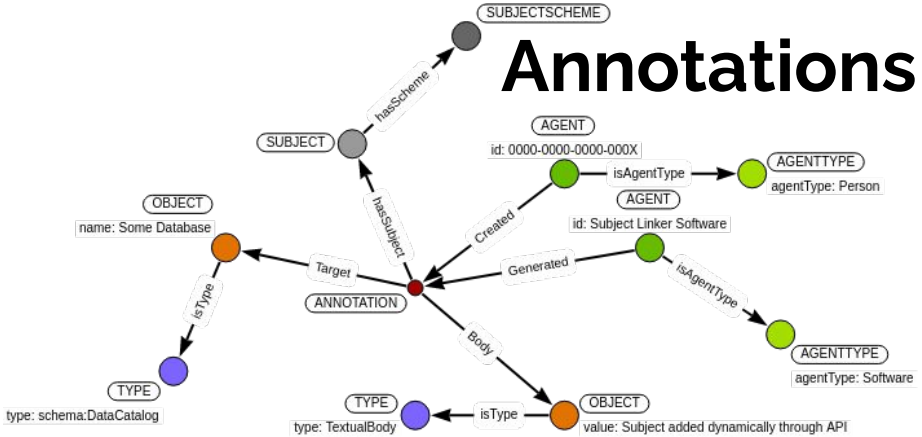
Extended using defined vocabularies (schema.org, DataCite, CrossRef, NSF)

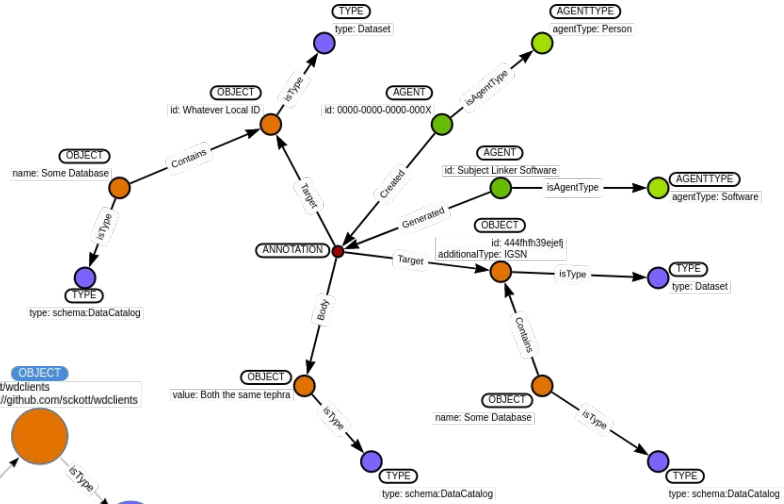
- Annotations add information (“this dataset is cool!”)
- Annotations link records (“This GitHub repo is mentioned in this paper”)

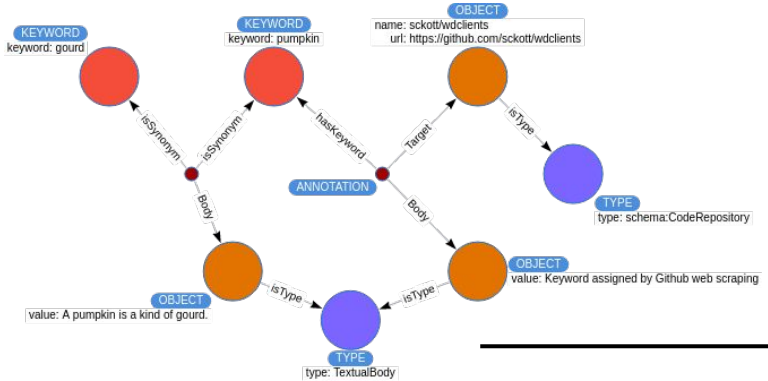
At its Heart, Annotations

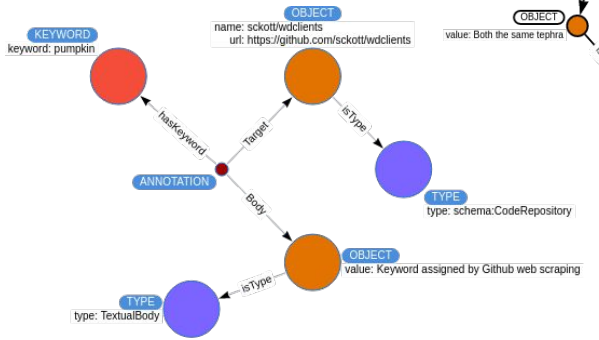


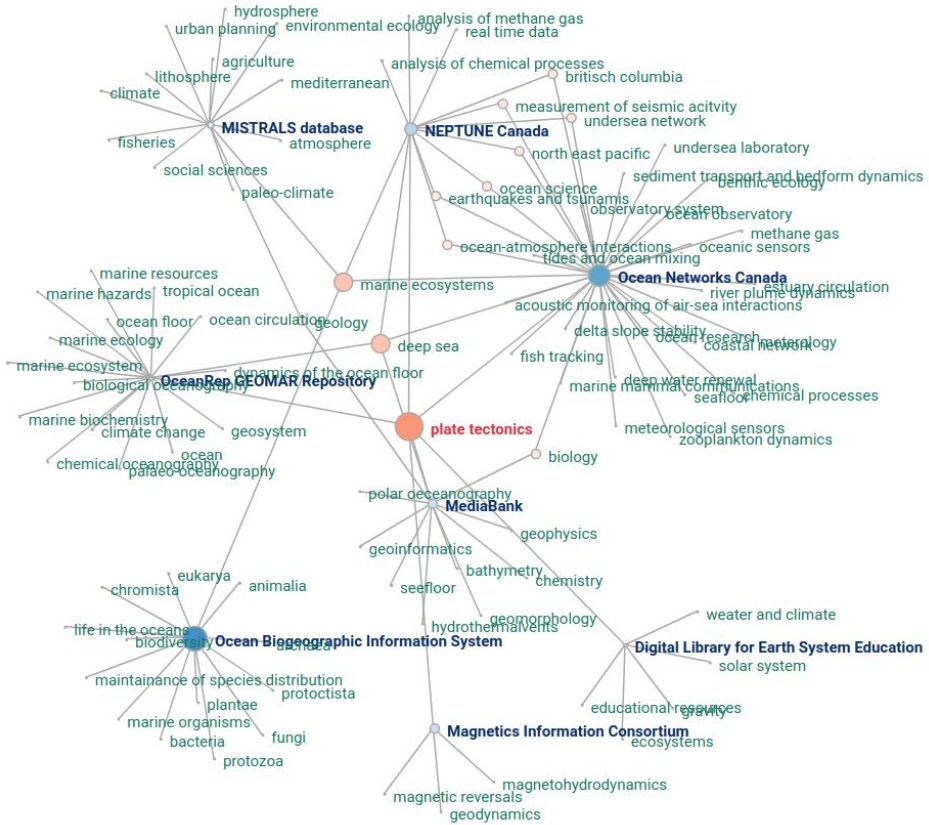
Annotations are Flexible











Linked Annotations Build Networks

- 300,000 Annotations
- 43,000,000 object connections
- 467,000 Scientific Grants
- 160,000 Research Articles
- 2,500 Databases
- 74,000 Code Repositories

Networks for Knowledge

1

Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.

Networks for Knowledge

1

Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

DataCite
FIND, ACCESS, AND REUSE DATA

Networks for Knowledge

1

Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

DataCite
FIND, ACCESS, AND REUSE DATA

Crossref

Networks for Knowledge

1

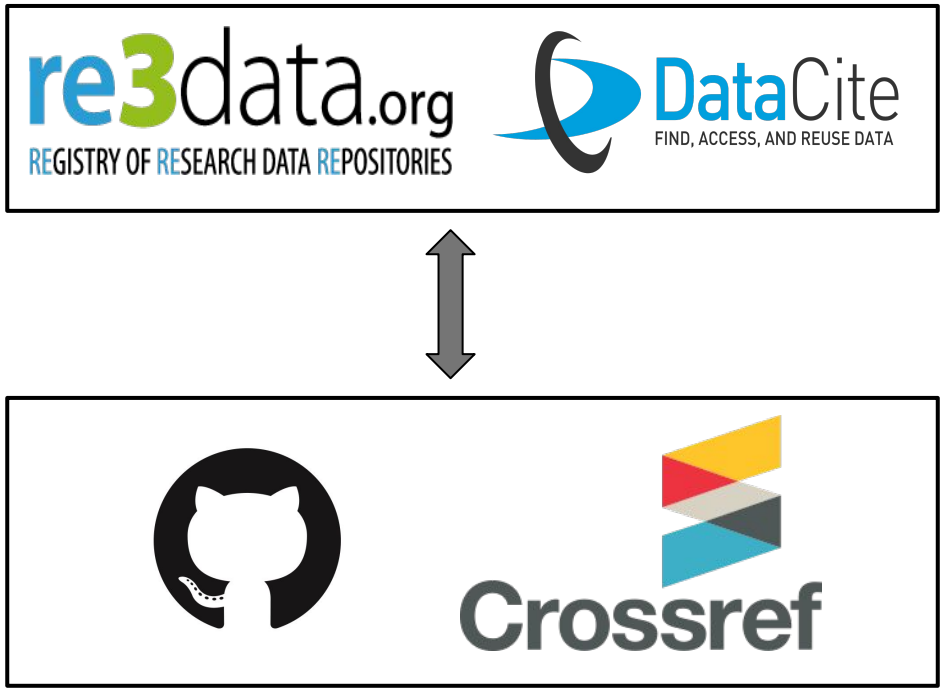
Finding Appropriate Data

Knowledge of disciplinary archives (e.g., Neotoma, OpenContext).

2

Using Appropriate Methods

Time series analysis, data processing, R packages.



Discovering Links

AdamWilsonLabEDU/geo503-2018-finalproject-nathandubinin

final_project.R

```
9   library(grid)
10  library(neotoma)
11  library(analogue)
12
13  # Info from Neotoma Explorer (Data originally from European Pollen Database)
…
31  # Getting data using Neotoma's API for
32  # Lake Sambösjön
33  sam <- get_site(sitename = 'Lake Sambösjön')
34  sam_pollen=get_dataset(sam)
```

R Showing the top six matches Last indexed on 14 Dec 2018

Neotoma
R package

xDeepDive (geoDeepDive)

13,415,017 documents

47,417 added this month

14,140 added this week

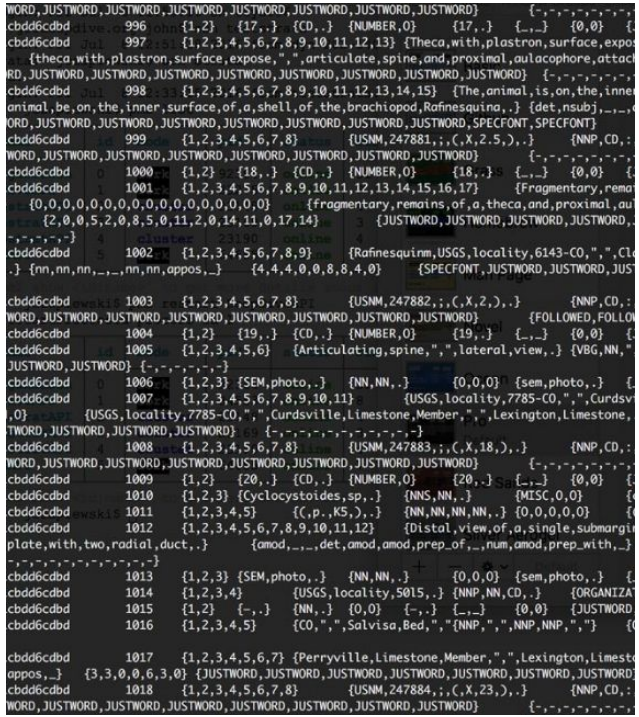
2,539 added in the last 24 hours

https://geodeepdive.org/

xDeepDive

Search & Tagging with:

- Full Text
- Regular Expression
- NLP Models



Shanan E. Peters
Dean L. Morgridge Professor

Sedimentary Geology and Paleobiology, Geoinformatics

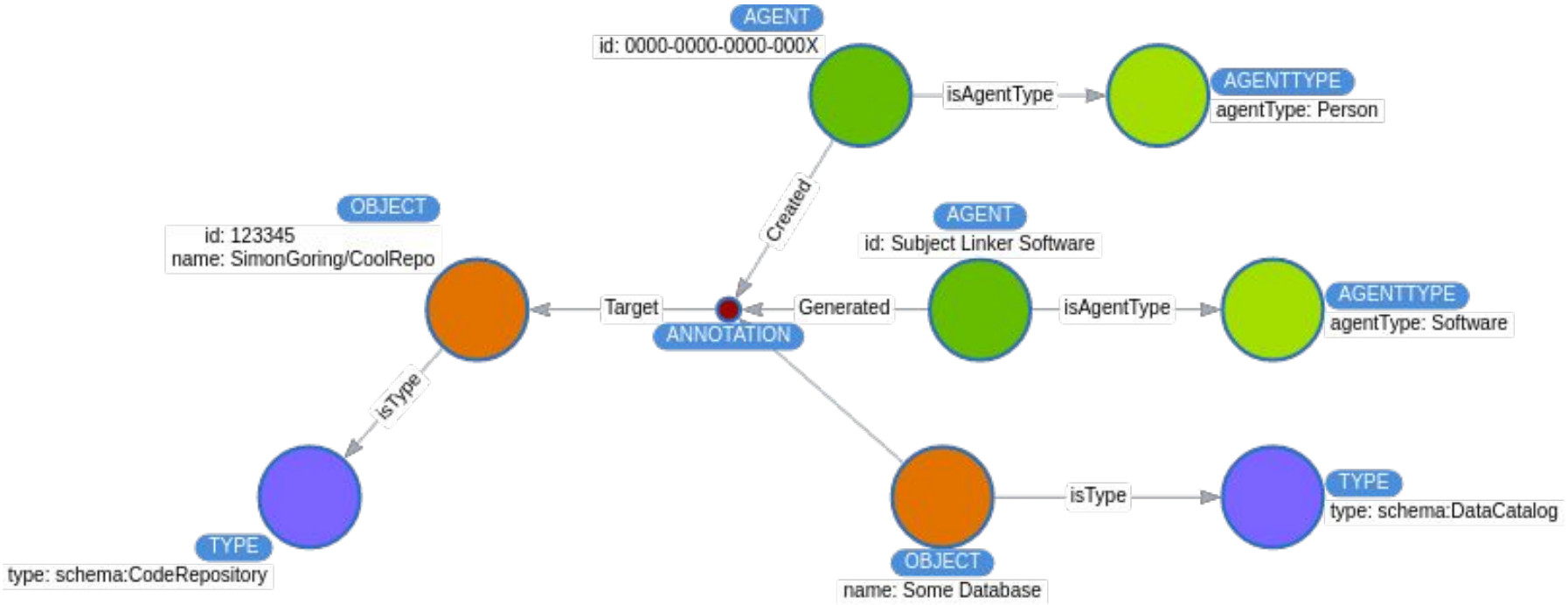
Crossref

EarthCube
TRANSFORMING GEOSCIENCES RESEARCH

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

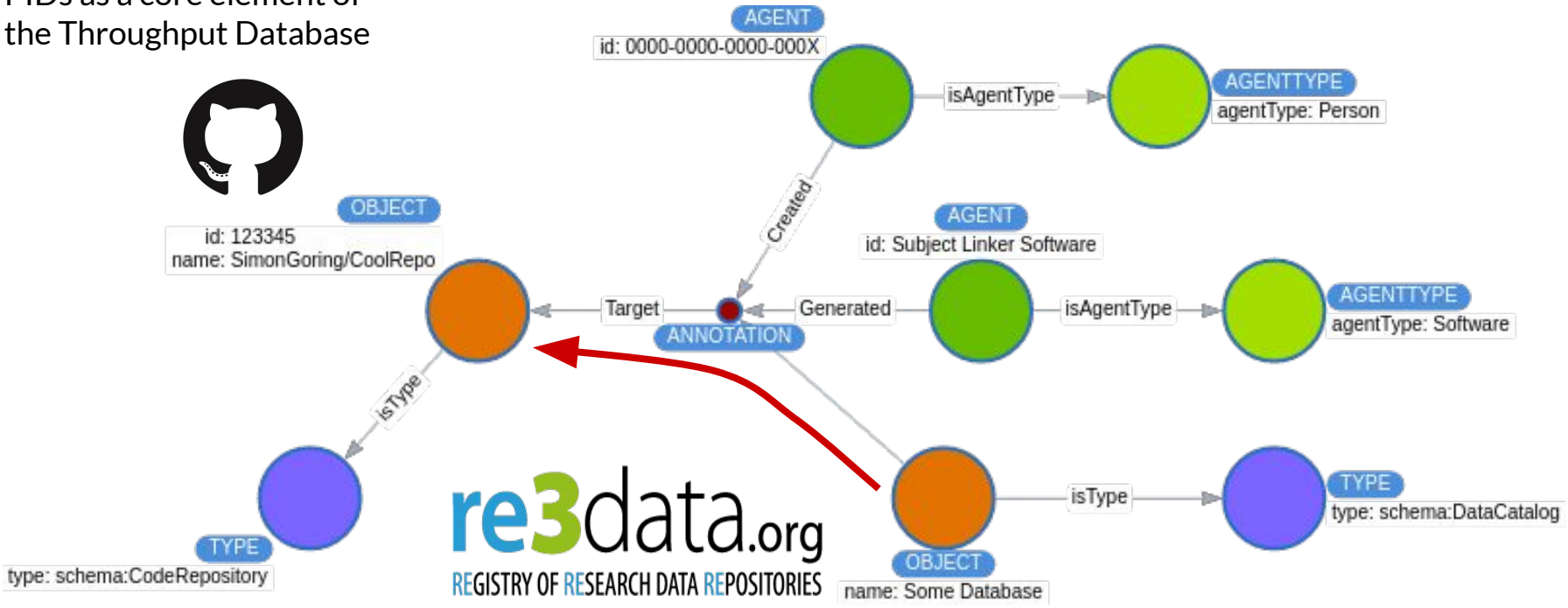
NSF

DB to GitHub Links



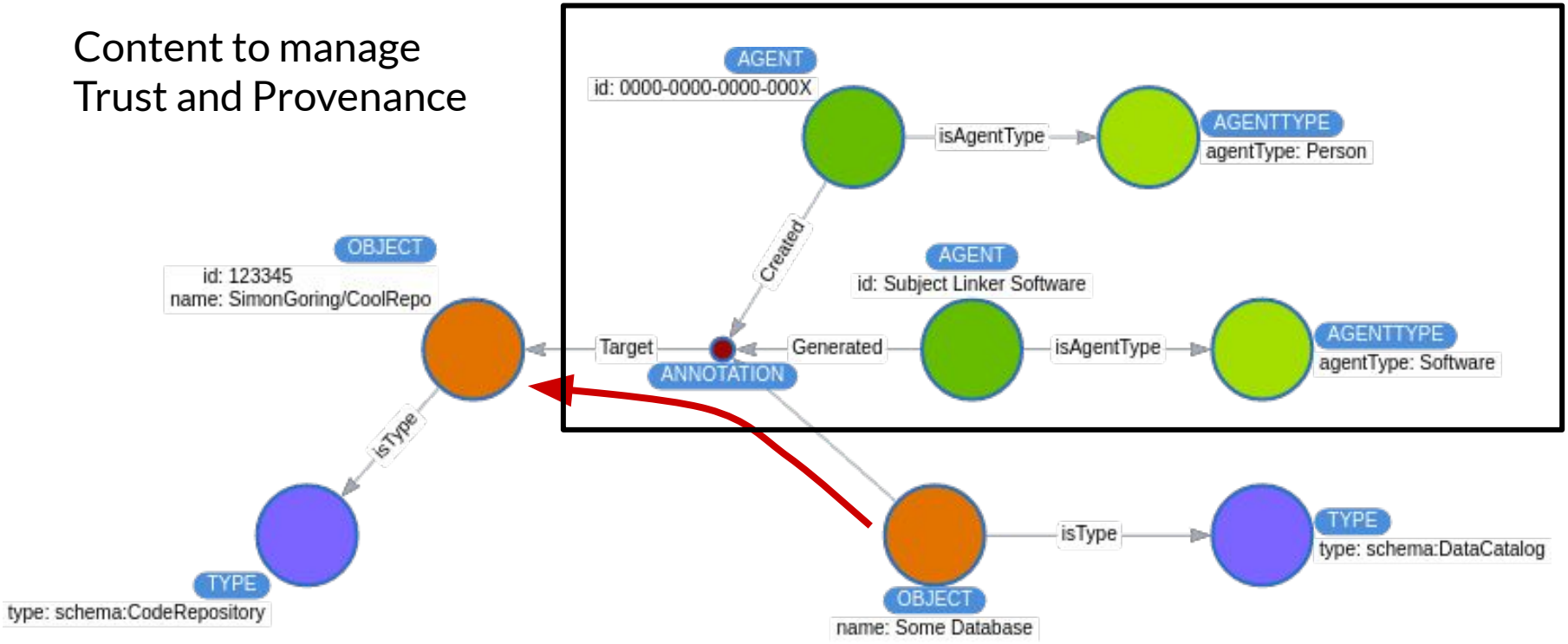
DB to GitHub Links

Use of External APIs and PIDs as a core element of the Throughput Database



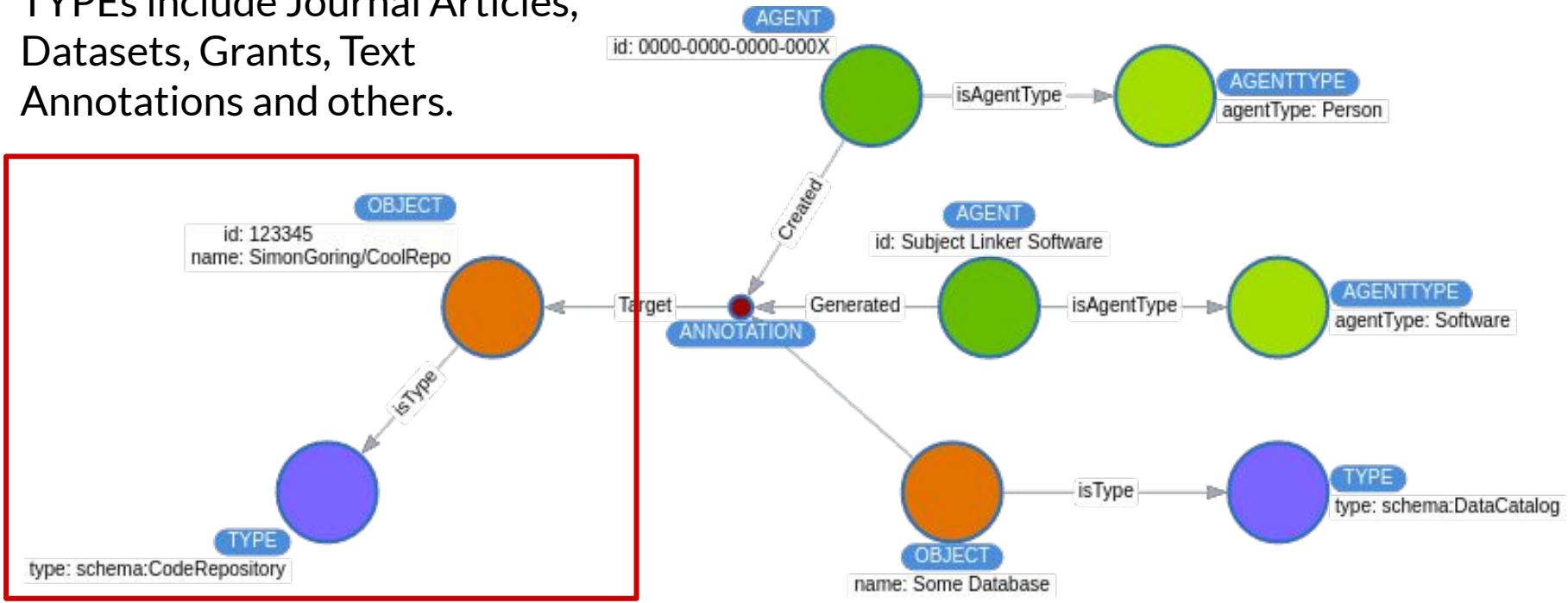
DB to GitHub Links

Content to manage
Trust and Provenance

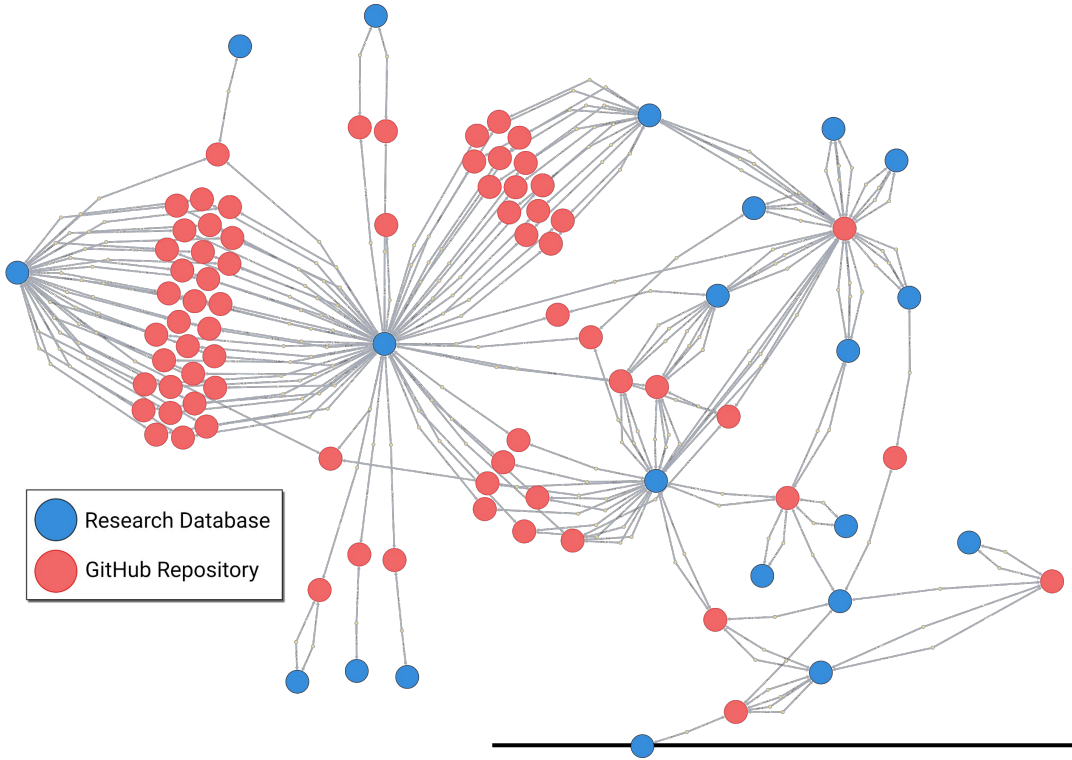


Model Is Extensible

TYPEs include Journal Articles, Datasets, Grants, Text Annotations and others.

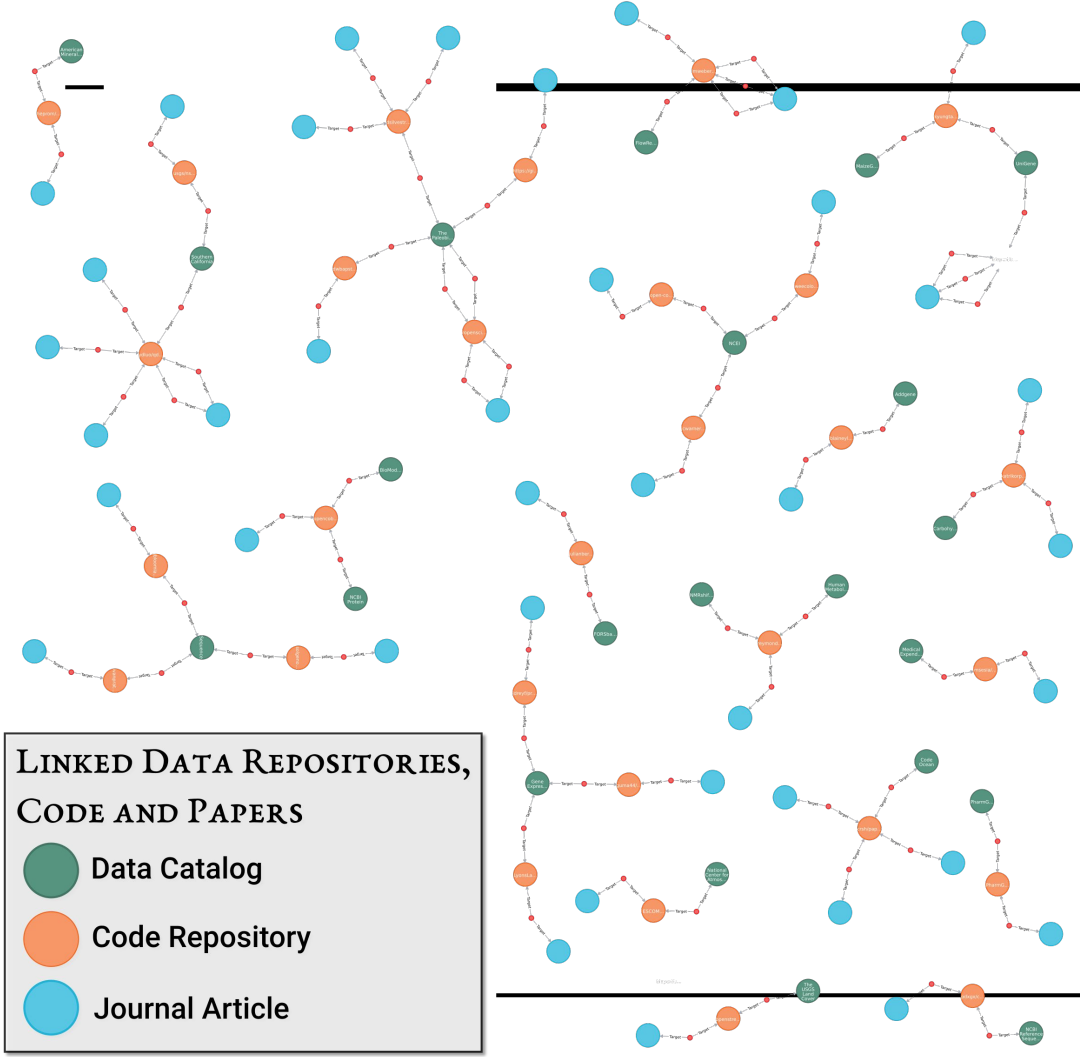


Links Across Data Resources



Patterns of data alignment

Variety often poses one of the greatest challenges in Big Data analysis. Are these repositories sources of information on how best to undertake data alignment?



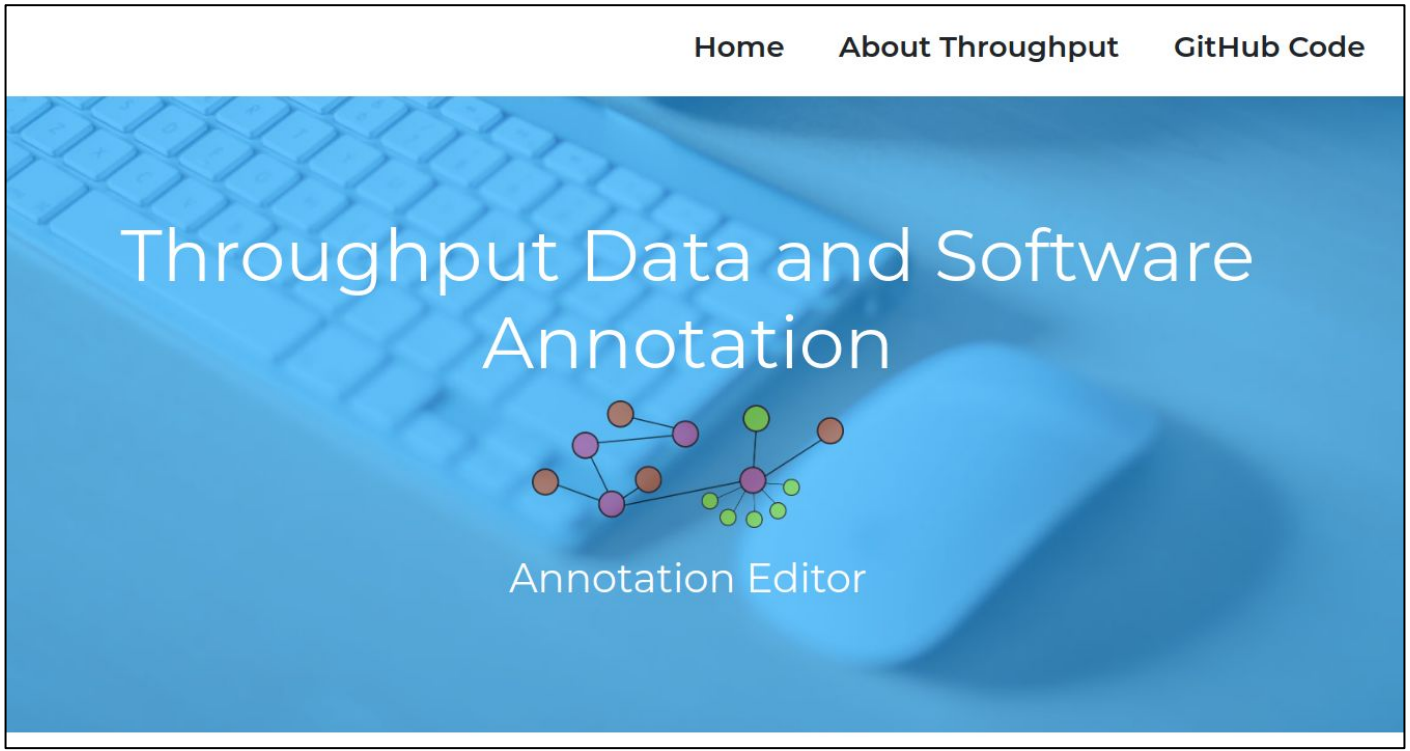
Data Discovery

Tracking Data Use

Understanding Best Practices

Retrospective Credit & Linking

Data Discovery & Use



Challenges In Metadata Recovery

README Files (https://github.com/throughput-ec/Template)

Proper citation & identification of source data

Optical Character Recognition issues in legacy publications

Lack of standards for Jupyter/RMarkdown documents (https://github.com/earthcube/NotebookTemplates)

Challenges in Code Discovery

README Files (https://github.com/throughput-ec/Template)

Proper citation & identification of source data

Optical Character Recognition issues in legacy publications

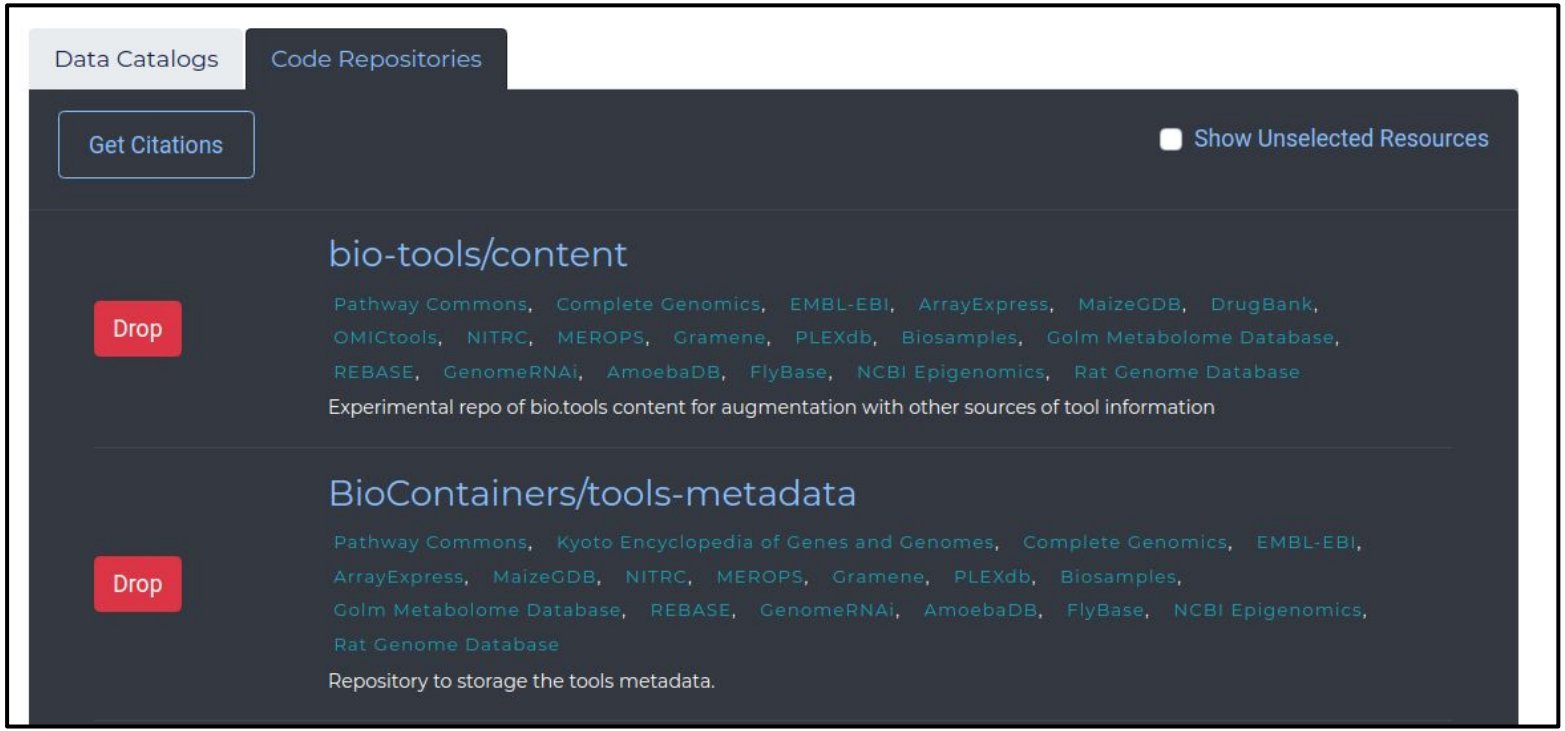
Lack of standards for Jupyter/RMarkdown documents (https://github.com/earthcube/NotebookTemplates)

Repository Variety

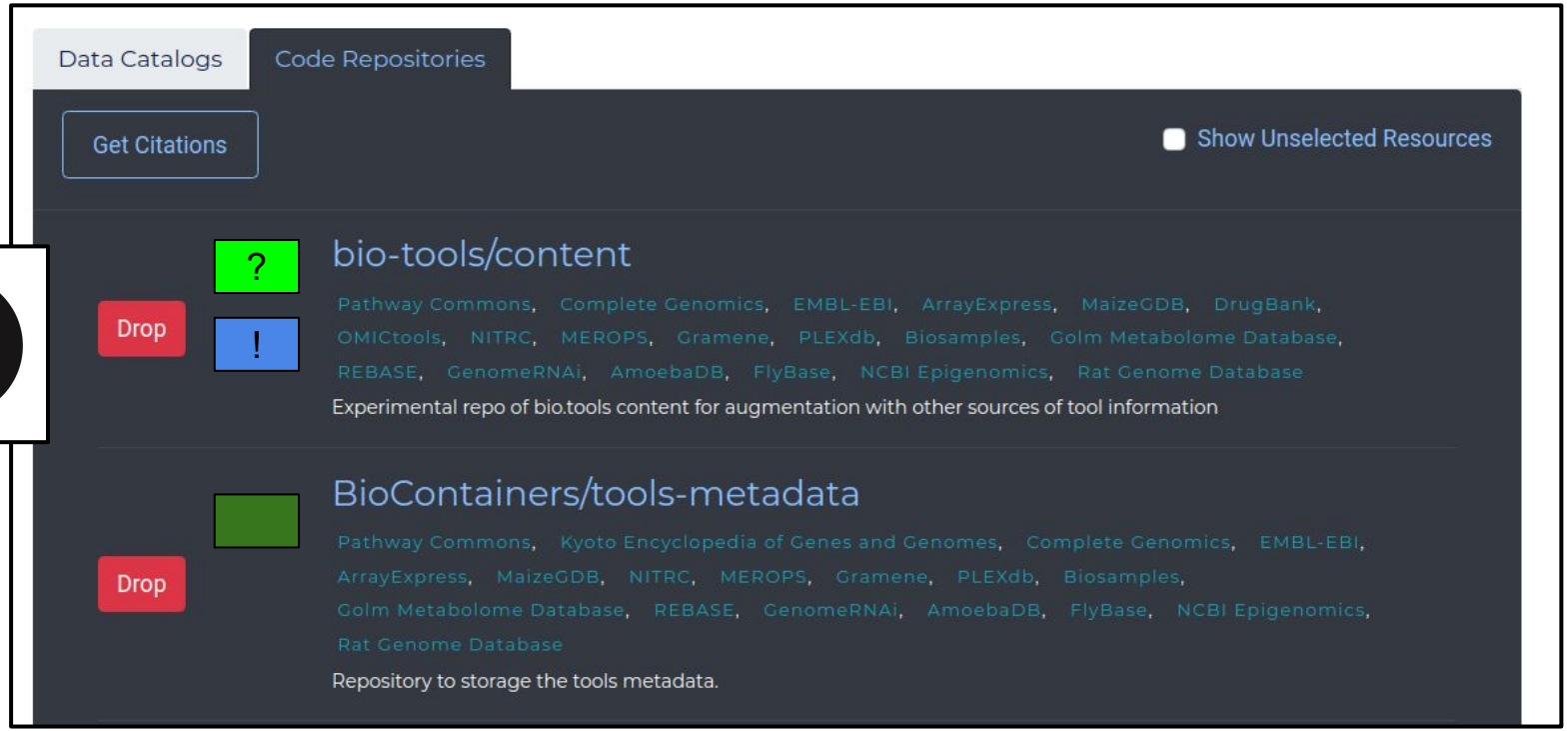
GitHub Repository Typology

Throughput
Educational: Elements of this repository serve to house educational and instructional resources. This may include: Books, assignments, class notes, and tutorials. This also includes manuals and training materials for the use of software. This does NOT include the results of class projects.
Analysis: This repository contains the data and code used as a primary analysis for some sort of research project. These are custom data analysis pipelines, not meant to be generally reusable. This category includes use and reuse of data. This includes: comparative reuses, original research studies, meta-analyses, statistical method development, and reproducibility tests.
Software development: Elements of this repository serve to build freestanding tools of any sort including libraries, plugins, frameworks, etc. This includes of use data to pilot tools. Includes calls to database API or library or other computational interface. Includes calls to R libraries
Storage: Repository stores copies of data from research databases.
Miscellaneous: This category is for repositories that are out of scope of this typology. This includes:
Scraping database registries: Repository contains lists of research databases and associated metadata.
Articles referencing database: Repository contains articles that link to research databases.
Informational link to database: Links to research database's homepage or another informational page. Not to a dataset.
Can't categorize/not enough information
Github repository no longer accessible: Github repository gives a 404 error.
Research database no longer accessible: Research database linked from GitHub repository gives a 404 error.

ML Classification and Tagging



ML Classification and Tagging



Active Development & Research

Implementing end-user annotation widget for data resources

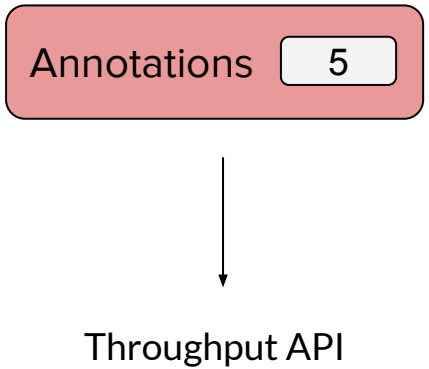
Characterizing patterns of data use and reuse in the geosciences

Feature extraction & metadata generation from publication

Data repository recommendation system

Continued integration of legacy & informal knowledge into the research data graph

Active Development & Research



Implementing end-user annotation widget for data resources

Characterizing patterns of data use and reuse in the geosciences

Feature extraction & metadata generation from publication

Data repository recommendation system

Continued integration of legacy & informal knowledge into the research data graph

Active Development & Research

Implementing end-user annotation widget for data resources

Characterizing patterns of data use and reuse in the geosciences

Feature extraction & metadata generation from publication

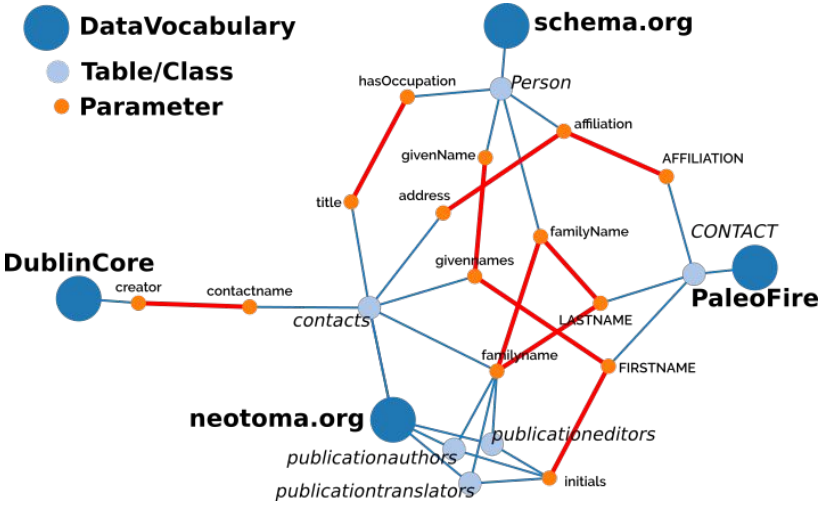
Data repository recommendation system

Continued integration of legacy & informal knowledge into the research data graph

Characterizing patterns of data use

Volcanic Eruptions and Their Repose, Unrest, Precursors, and Timing

National Academies of Sciences, Engineering, and Medicine **2017**. Washington, DC: The National Academies Press. https://doi.org/10.17226/24650



Active Development & Research

Implementing end-user annotation widget for data resources

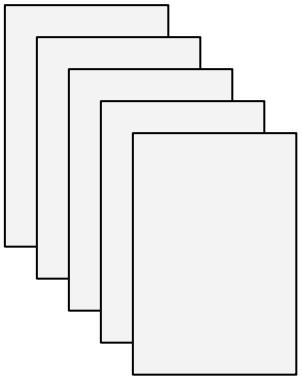
Characterizing patterns of data use and reuse in the geosciences

Feature extraction & metadata generation from publication

Data repository recommendation system

Continued integration of legacy & informal knowledge into the research data graph

Data Pipeline

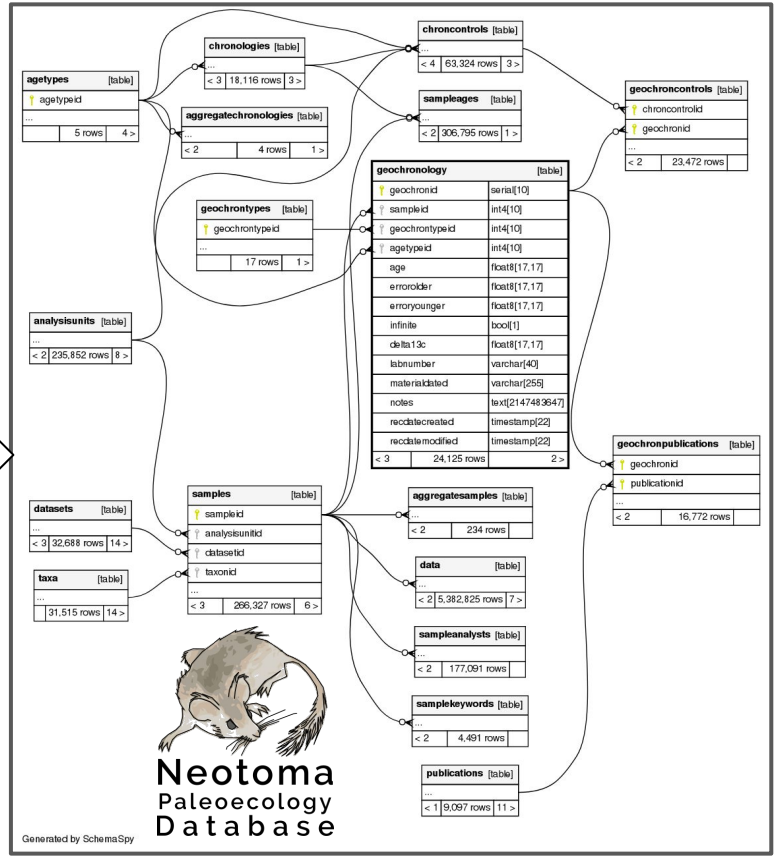


bit.ly/emptyspacegdd

regex tagging

ML Prediction

Socorro Dominguez



Active Development & Research

Implementing end-user annotation widget for data resources

Characterizing patterns of data use and reuse in the geosciences

Feature extraction & metadata generation from publication

Data repository recommendation system

Continued integration of legacy & informal knowledge into the research data graph

Thanks!

Technical Notes

1. License - MIT (http://github.com/throughput-ec/throughputdb)
2. Data Standards Implemented
    a. W3C Annotation (https://github.com/throughput-ec/throughputdb/tree/master/cypher_anno_examples)
    b. Schema.org
    c. existing (parent) metadata formats
3. Notebooks
    a. DeepDive work using Jupyter notebooks & Dockerization (https://github.com/throughput-ec/UnacquiredSites)
    b. DB/API/App all reproducible (http://github.com/throughput-ec), Dockerization to follow.
4. Fully version controlled (all work at http://github.com/throughput-ec)
5. Services
    a. REST API using JSON (JSON-LD to follow), Documented using OpenAPI v3.0 (developed but buggy: https://throughputdb.com/api-docs)
6. Data
    a. Snapshots deposited in Figshare (http://doi.org/10.6084/m9.figshare.12731138)