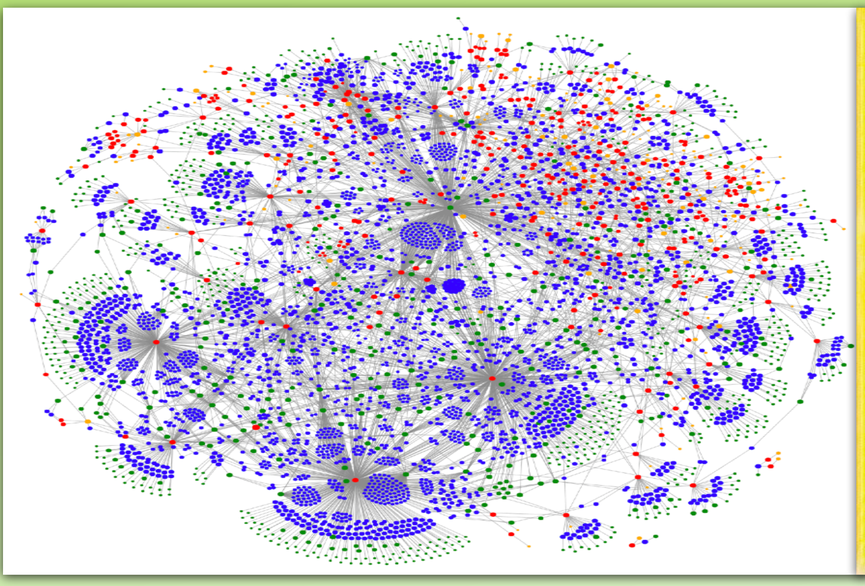
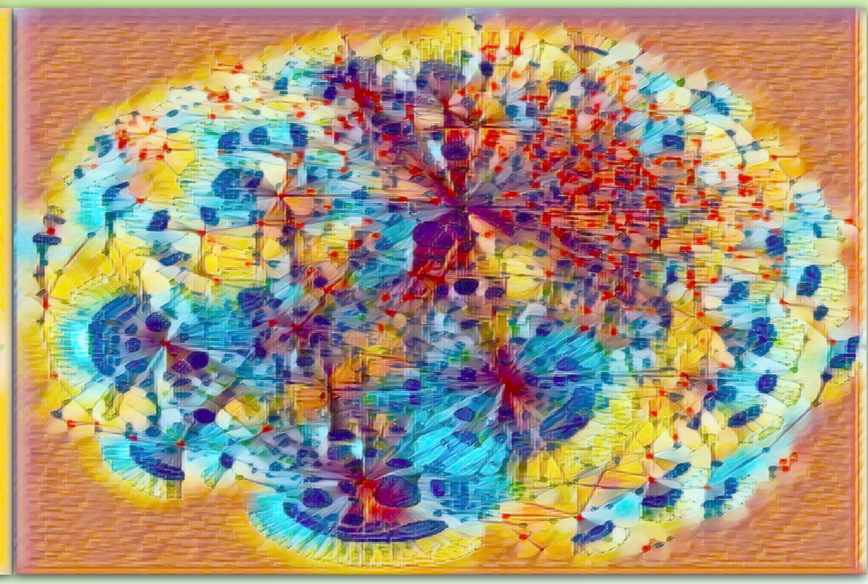
Rich Context:

Rich Search and Discovery for Scholarly Datasets

Paco Nathan @pacoid

derwen.ai



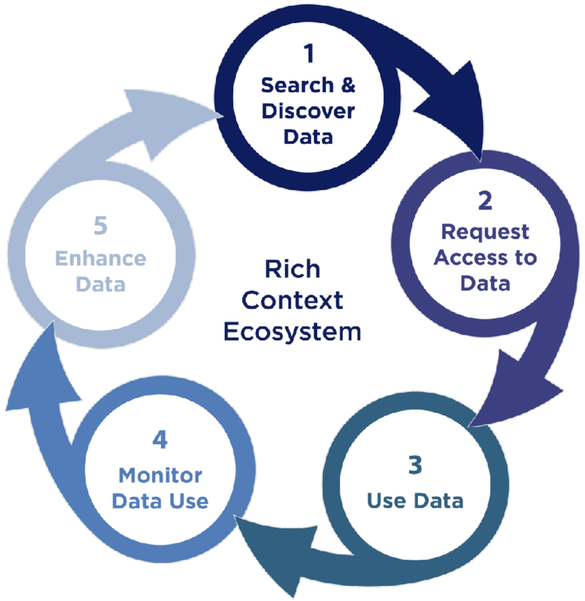


Administrative Data Research Facility

Coleridge Initiative

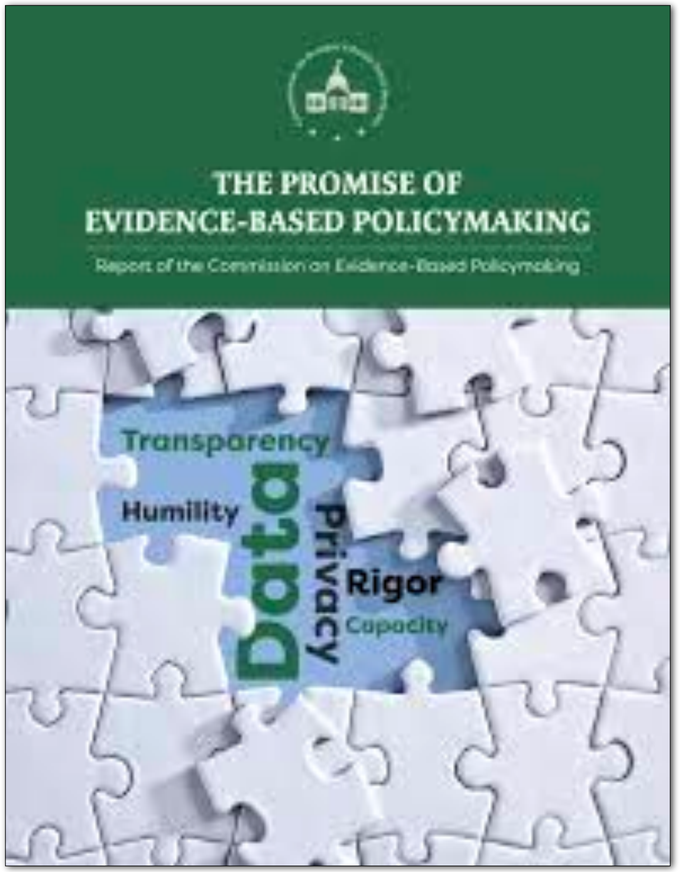
Julia Lane, et al. NYU Wagner

- FedRAMP-compliant **ADRF framework** on AWS GovCloud: “public agency capacity to accelerate the effective use of new datasets”
- for research projects using cross-agency sensitive data, in US and EU – **now in use by 30+ agencies**
- cited as the first federal example of Secure Access to Confidential Data in the final report of the Commission on **Evidence-Based Policymaking**
- augments **Data Stewardship** practices; collaboration with Project Jupyter on the related **data gov features**



Related Federal Directives

- **Foundations for Evidence-based Policymaking Act** (2018)
- **Information Quality Act** (2001)
- **NIH Strategic Plan for Data Science** (2018)
- **US federal data strategy**
- **Year-1 Action Plan** (2019)



See also:

"Evidence-based decision making: What DOE, USDA and others are learning"
Wyatt Kash
FedScoop (2019-06-28)

Training

Data producer

Data steward

Metadata
Data

Usage
Feedback

Access
Workflows
Monitoring

Reporting

Data user

Metadata

Data analysis
Code
Collaboration

Rich Context

Links metadata, code, tools, publications

Data Stewardship

Approval workflow, monitoring, reporting

Collaboration

Interactive chat and code sharing

Workspace and tools

python™ jupyter R

Data in cloud

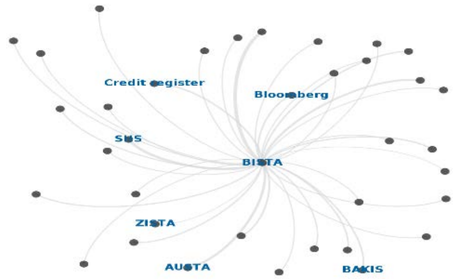
Alternative: local servers

Security

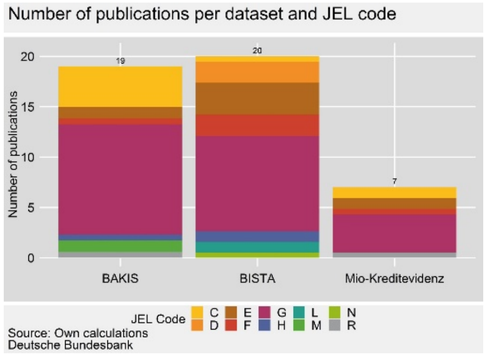
FedRAMP security certified

Research Data Centers

To date, we successfully launched a machine learning competition and obtained algorithms to extract used datasets from publications. Based on the results, we implemented two prototypes that will serve as building blocks for a unified system.



Recommend data to researchers (“*based on your interest, you might also like this data*”).



Data impact factor (*“This dataset generates most research/ societal / policy value”)*

The two prototypes already create value by enabling more optimal data usage and by supporting effective resource allocation into valuable data that generates results.

DEUTSCHE BUNDESBANK
EUROSYSTEM

– Hendrik Doll, Deutsche Bundesbank

The bigger picture

Our vision is a system, where researchers and analysts apply for confidential data access, get data recommendations, have a secure remote digital workspace, and provide feedback on data. Such a system relies on three non-traditional data sources: Data usage information, implicit knowledge in researchers’ heads (incentives for sharing), and structured administer metadata (*annodata*) to automatically govern data access.

Research Data Centers

Factsheet: Building an integrated data access system for empirical research

Hendrik Doll, Stefan Bender, Jannick Blaschke, Christian Hirsch, Christian Resch[1]

Federal Reserve Board, Washington, D.C., October 1-2, 2019

Our idea

Empirical research and evidence-based policy decisions increasingly rely on microdata. Research publications are well referenced and findable, however structured information on data usage is less available. The current project aspires to change this by building a data-centric ecosystem with rich context and a community around microdata.

Approach: Technical and Human

Technical

- Create secure environment where data providers can share their data across agency and jurisdictional lines
- Census and USDA Authorization to Operate; HHS in process

Operational

- Link disparate data
- Analyze data

Legal & Practical

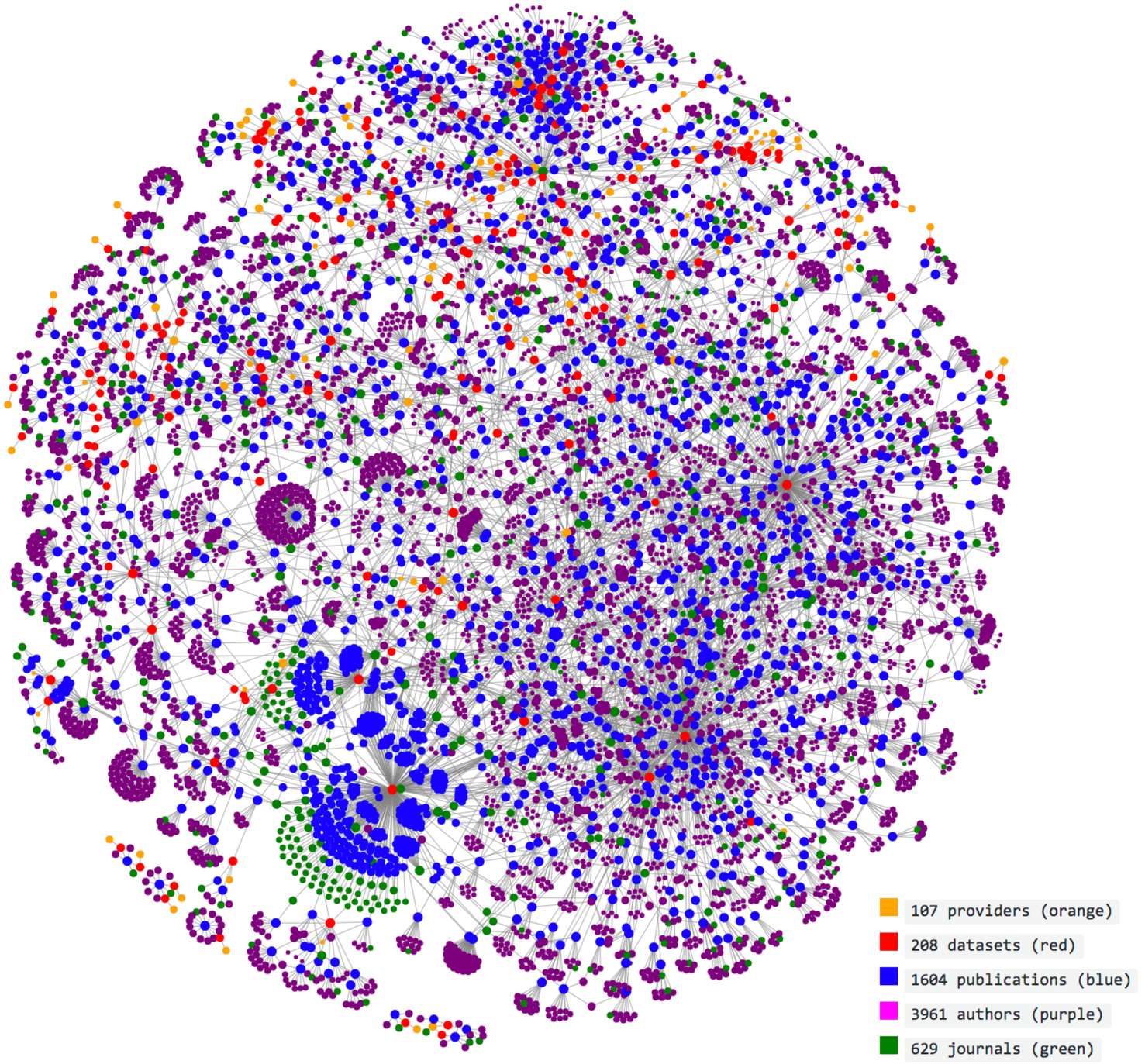
- Document value associated with the data linkage
  - Consistent with the agency mission
  - Useful enough to engage decision-makers

Recommenders for researchers/analysts

- Objective: provide better means of *search and discovery* for social science researchers and agency analysts.
- Collect workflow telemetry and query logs to augment the graph.
- Currently developing recommender systems based on the graph.
- This accelerates research and also assists training (e.g., onboarding agency analysts).
- Near-term goal: identify people with specific expertise.
- Long-term goal: learn workflow configurations to support AutoML meta-learning.

Rich Context

- Focus: identify findings and results consistent with the agency mission, and document the value and use of the data for evidence-based decision making
- Funded by Schmidt Futures, Alfred P Sloan Foundation, Overdeck Family Foundation
- Partnering with Bundesbank, USDA, etc.
- Collaboration with SAGE Pub, RePEc, ResearchGate, Digital Science, etc.



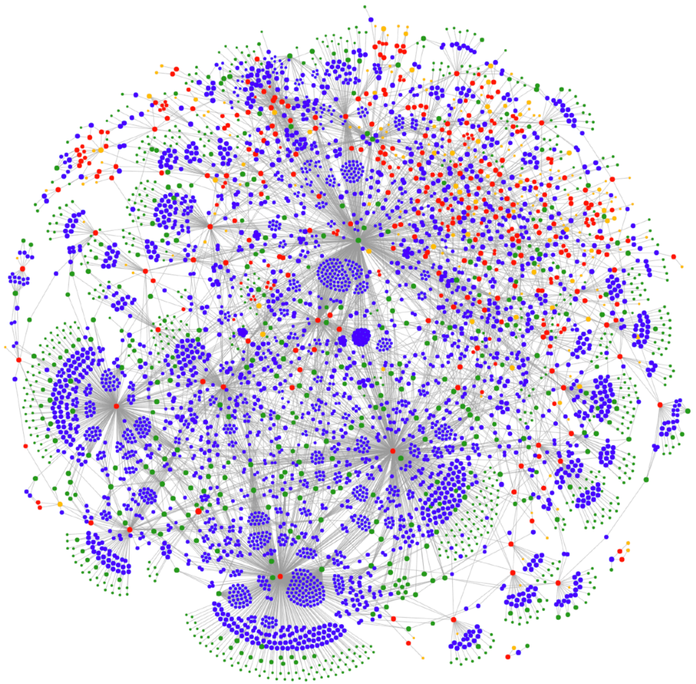
Rich Context

- Focus: identify findings and results consistent with the agency mission, and document the value and use of the data for evidence-bas making
- Funded by Schr Alfred P Sloan F Overdeck Famil
- Partnering with Bundesbank, USDA, etc.
- Collaboration with SAGE Pub, RePEc, ResearchGate, Digital Science, etc.

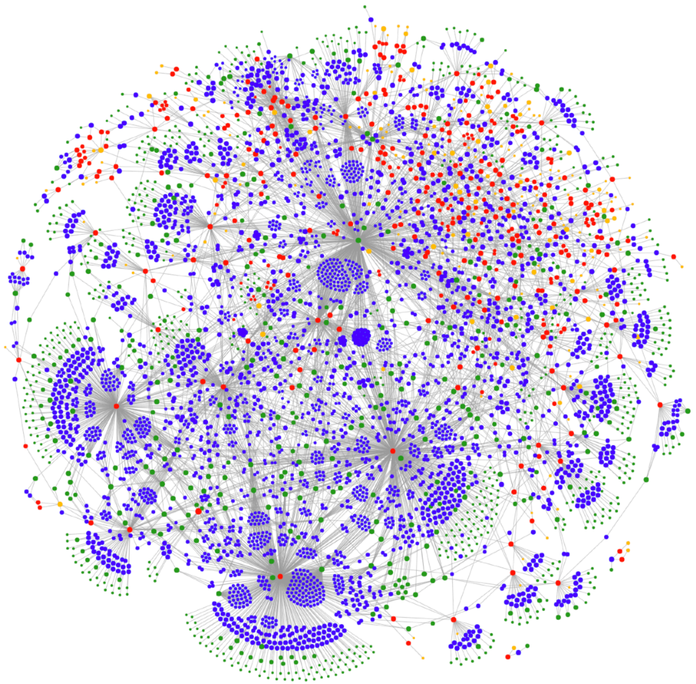
Challenges that empirical researchers face: for a given dataset, find out **who** has worked with the data before, **what methods and code** were used, and **what results** were produced.

Knowledge Graph – why?

- Allow flexibility for metadata representation
- Measure metadata quality
- Prepare features for ML models
- Build recommenders for *experts*, *topics*, *tools*, etc.
- Engage the public with automated data inventories
- Recommend configurations to new analysts
- Identify which datasets get used with others
- Quantify impact of datasets on policy



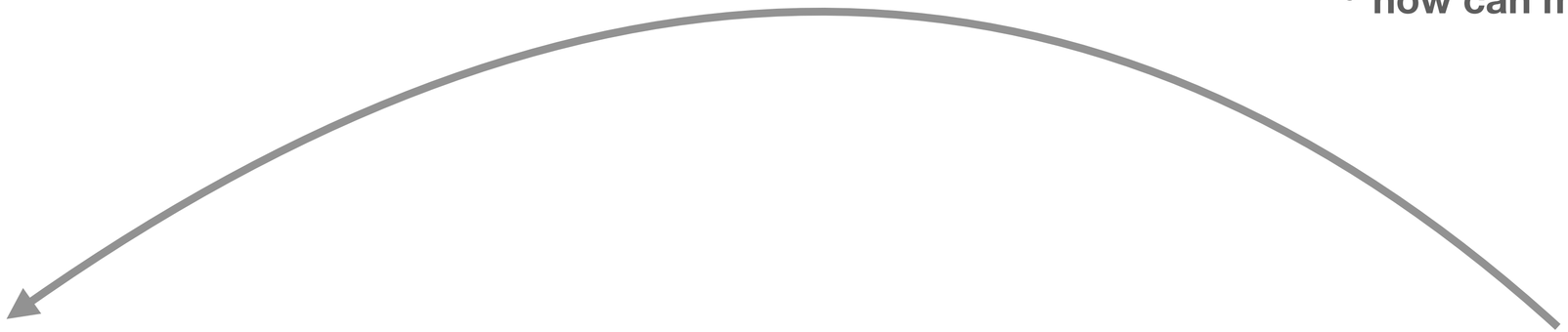
Knowledge Graph – how?



- Manual data entry and curation of linked data
- Use persistent identifiers whenever possible: DOI, ISSN, ROR, ORCID, etc.
- Leverage ML models to infer missing metadata
- Federate queries of discovery services APIs
- Suggest corrections for metadata errors
- Use HITL to build feedback loops that engage experts, and provide convenient means for manual override
- Identify errors by using unit tests, ontology axioms, graph analytics, etc.
- **Collaborate with agency libraries!**

KG process

- who are the expert people?
- which topics are emerging?
- how can methods be shared?



activities → **outputs** → **outcomes** → **impact**

curated datasets

research projects

published research

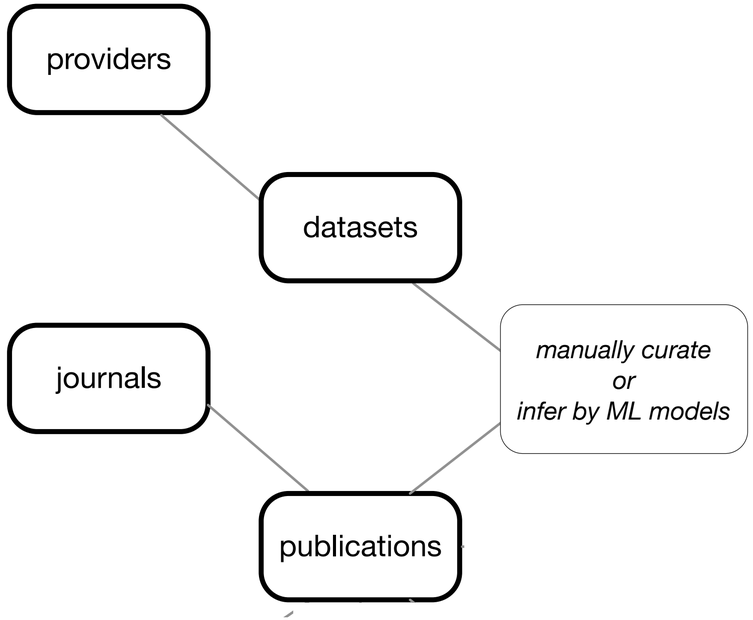
better science, government, education

ML models infer new metadata links

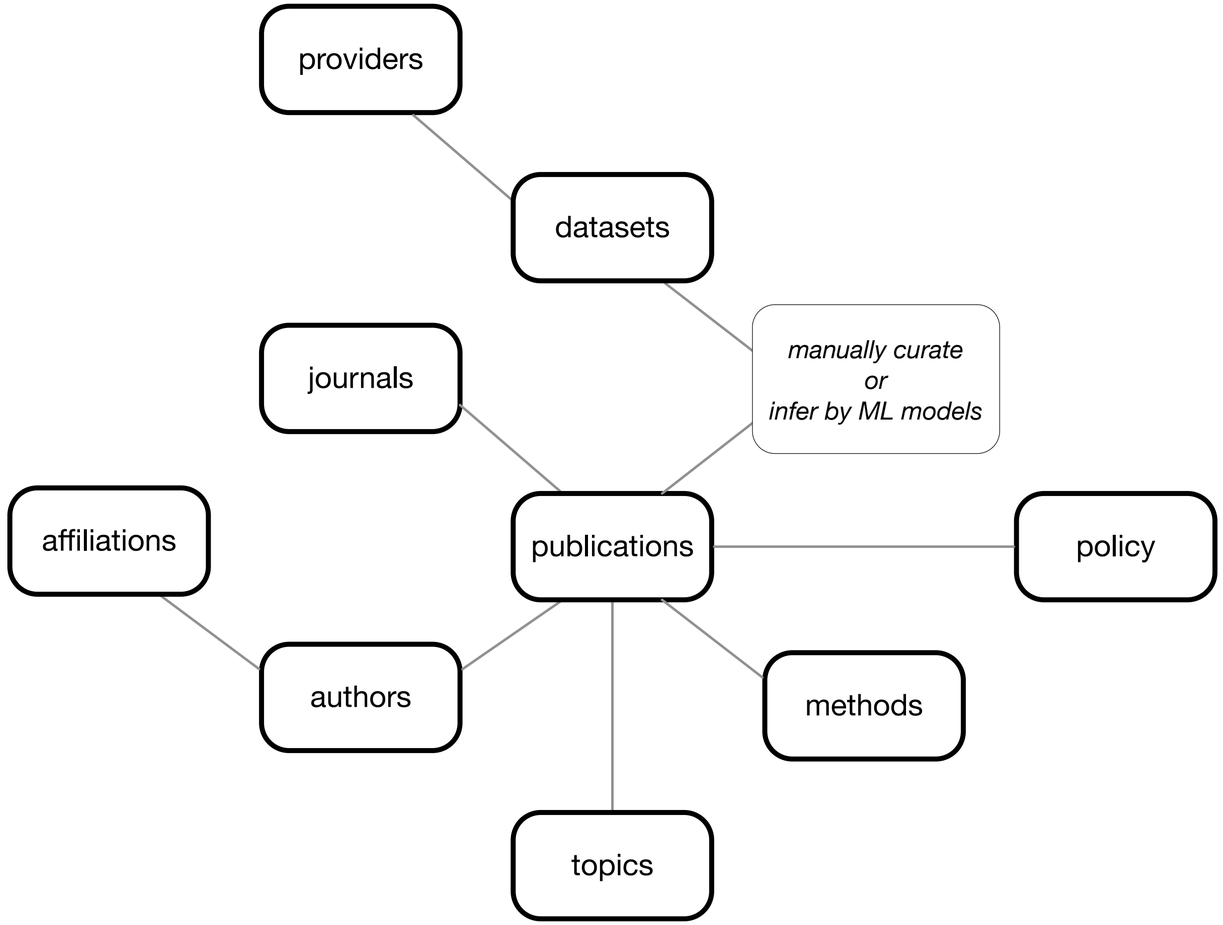
how do we track the linkage??

how do we measure these behaviors??

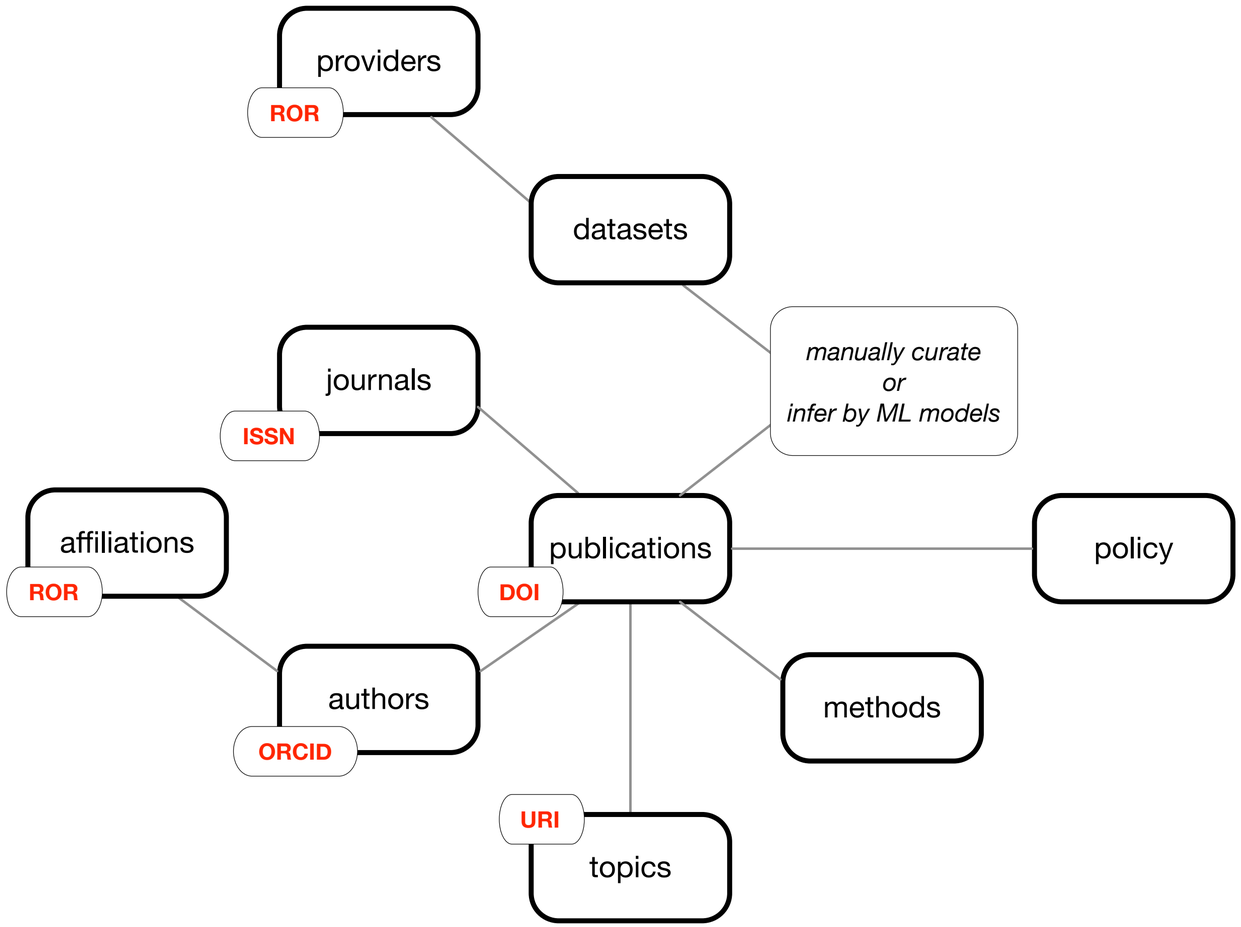
KG construction and representation



KG construction and representation

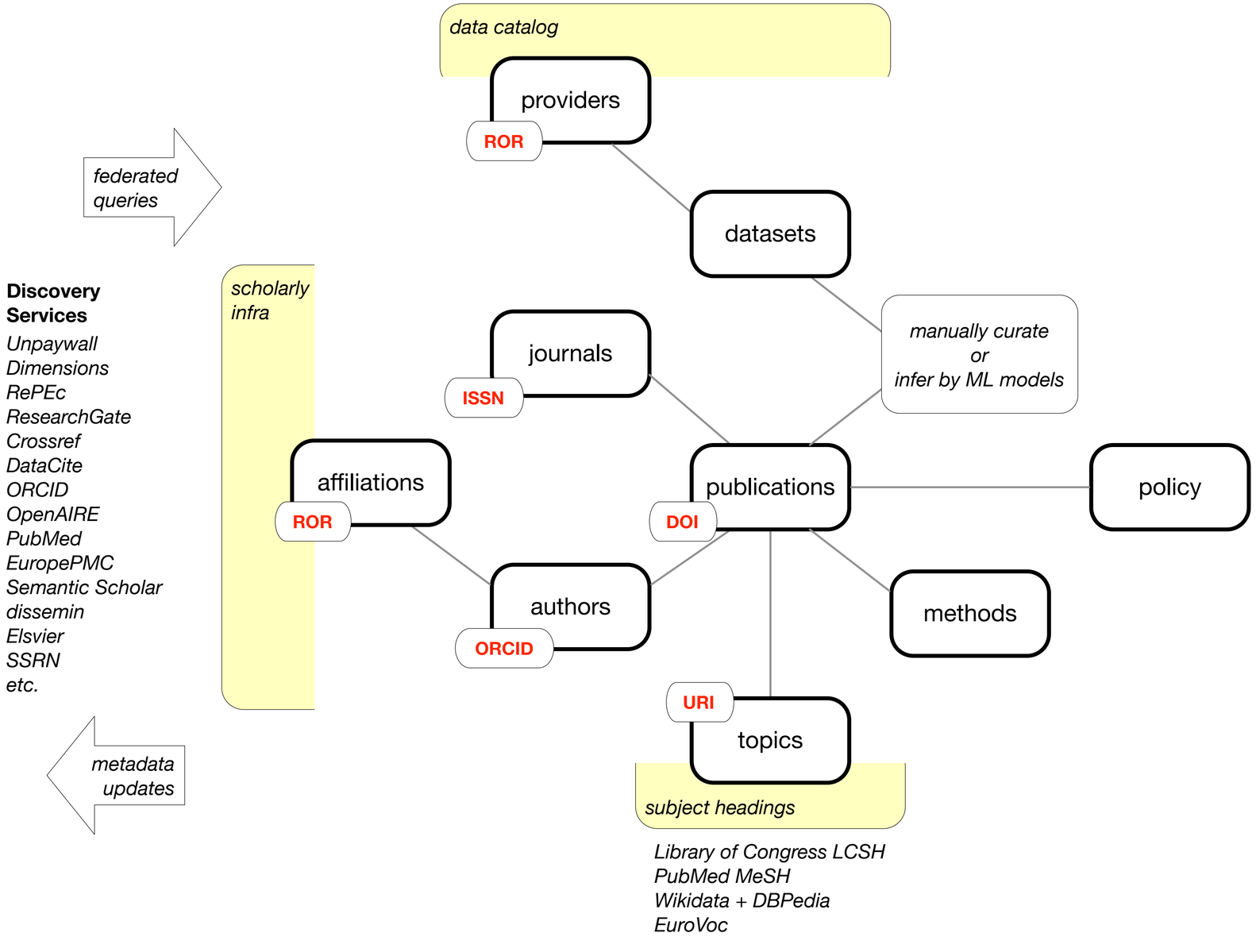


KG construction and representation

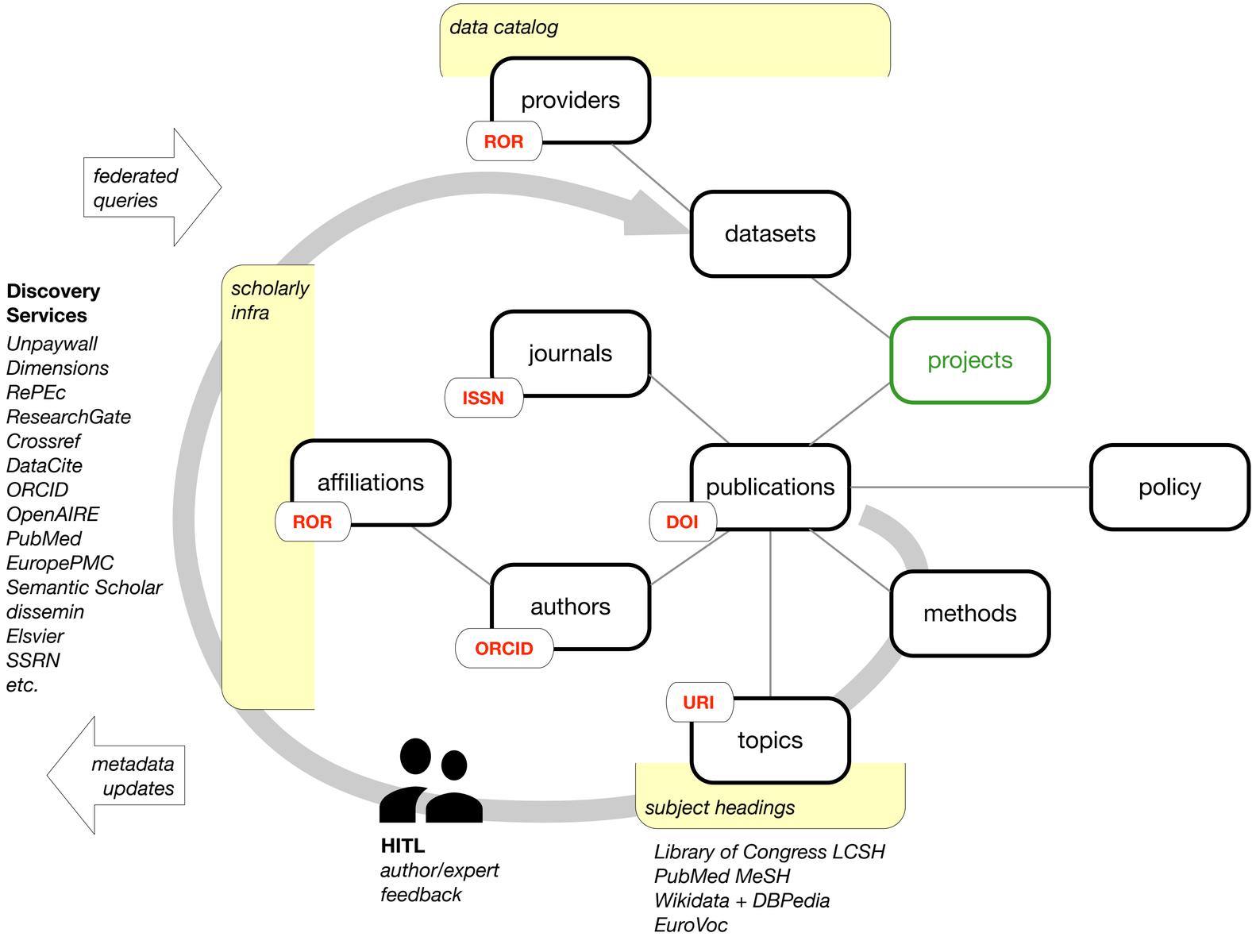


See github.com/Coleridge-Initiative/rclc/wiki/Corpus-Description

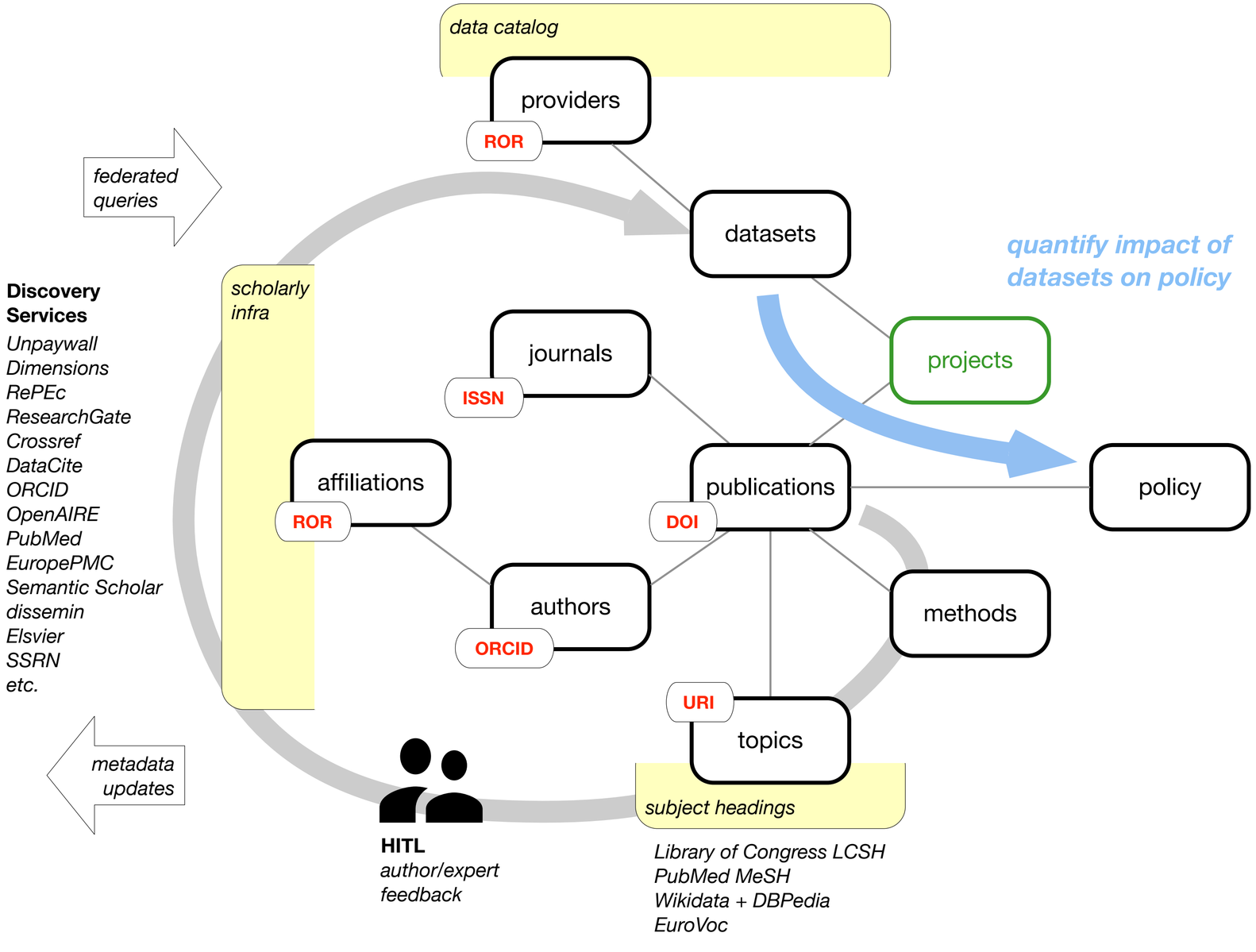
KG construction and representation



KG construction and representation



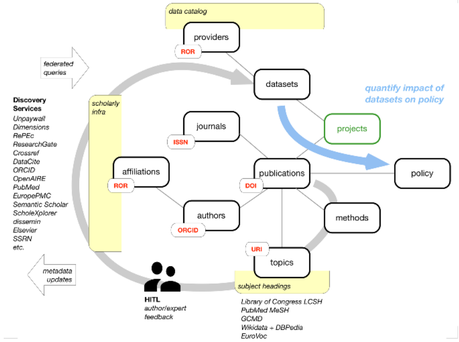
KG construction and representation



KG construction

As a **spaCy** pipeline extension, **PyTextRank** produces annotations for text documents: linked entities associated with other metadata. For example, given a corpus of research papers:

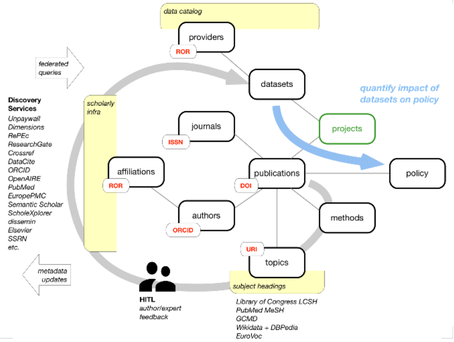
1. parse each PDF – e.g., with **Parsr**
2. identify top keyphrases from the “Conclusions” section
3. query **discovery services** for author list, journal, citation links, and so on (federating OpenAIRE, PubMed, Crossref, etc.)
4. use combined annotations as training data for vector embeddings
5. query trained model as an oracle: distances between an entity and its neighbors become edge weights for constructing a KG
6. prune the constructed KG, if needed – e.g., with a **disparity filter**



KG personas and use case priorities

User Persona:

- *auditor or congressional aide:*
  why/how was money spent collecting this dataset?
- *Jane Q. Public:*
  what are the uses for this dataset?
- *analysts/researchers:*
  research tool (our grad student researchers included)



Funders:

- supporting this effort to improve the process of scientific workflows

Open Source Projects

- **RCGraph** – Rich Context knowledge graph management
  github.com/Coleridge-Initiative/RCGraph
- **richcontext.scholapi** – federated discovery services and metadata exchange across scholarly infrastructure APIs
  pypi.org/project/richcontext-scholapi
- **RCLC** – ML leaderboard competition
  github.com/Coleridge-Initiative/rclc
- **adrf-onto** – controlled vocabulary for ADRF and Rich Context using OWL, SKOS, DCAT, PAV, CITO, FaBiO, etc.
  github.com/Coleridge-Initiative/adrf-onto

See also:

“Machine Learning Highlights for Rich Context”

Funded additions to Project Jupyter

Make datasets and projects top-level constructs, support metadata exchange and privacy-preserving telemetry from notebook usage:

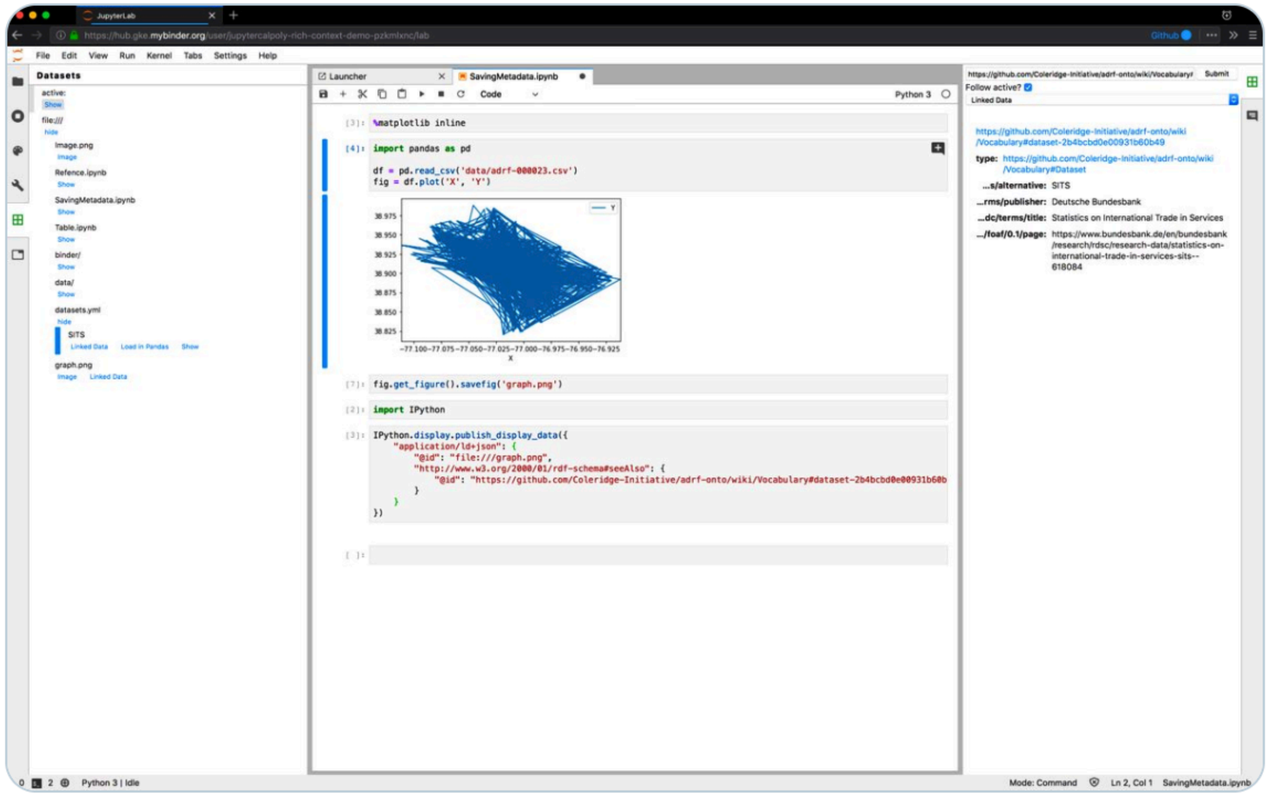
- JupyterLab **Commenting** and real-time collab similar to Google Docs
- JupyterLab **Data Explorer**: register datasets within research projects
- JupyterLab **Metadata Explorer**: browse metadata descriptions, get recommendations through knowledge graph inference (via extension)
- **Data Registry** (original proposal)
- **Telemetry** (privacy-preserving, reports usage)

Saul Shanabrook
@SShanabrook

twitter.com/SShanabrook/status/1182442214980501505

Replying to @choldgraf

Data Catalog Vocabulary and other related vocabularies are useful here. w3.org/TR/vocab-dcat/ We are building a way to explore metadata defined in JSON LD that uses these in JupyterLab github.com/jupyterlab/jup... cc @pacoid



12:46 AM · Oct 11, 2019 · Twitter Web App

ML Leaderboard Competition

github.com/Coleridge-Initiative/rclc

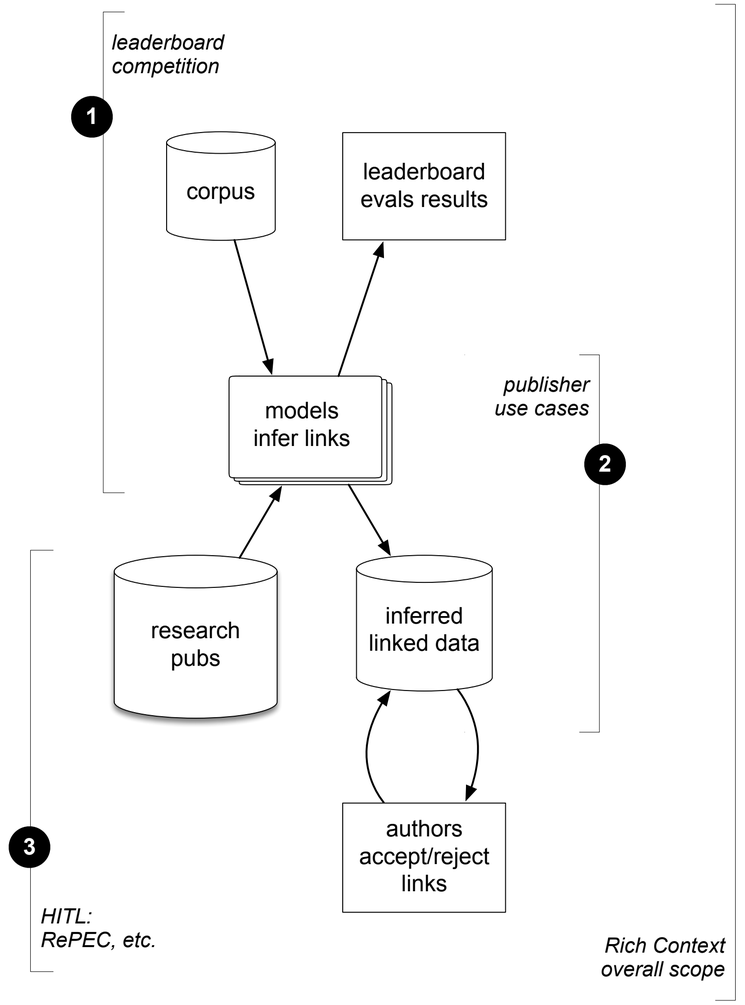
- update from RCC competition in 2018-2019
- ongoing ML leaderboards (similar to **NLP-progress**)
- open source, hosted on GitHub
- highly curated test sets, all open-access publications
- teams collaborate via GH issues on corpus data quality, etc.
- focus on *precision* for ML model evaluation

Current SOTA

source	precision	entry	code	paper	corpus	submitted	notes
LARC @philipskokoh	0.7836	ipynb	repo	RCC_1	v0.1.5	2019-09-26	RCLC baseline experiment using RCC_1 approach
KAIST @HaritzPuerto	0.6319	ipynb	repo	RCC_1	v0.1.5	2019-11-01	model trained a different dataset using DocumentQA and Ultra-Fine Entity Typing -- NB: this approach is able to identify new datasets

Human-in-the-loop

- active learning, aka“human-in-the-loop” – in progress via RePEc
- interact with authors to confirm metadata inferred by ML models
- feedback from experts improves the corpus metadata and the ML modeling



See also:

“Human-in-the-loop AI for scholarly infrastructure”

“New initiative to help with discovery of dataset use in scholarly work”, Christian Zimmerman

Agency Collaboration

- Develop reusable dataset discovery services, so that the public and researchers can find trustworthy, high-impact data
- Identify experts who have used the data and the associated research topics, associated analytical methods and tools, and related datasets
- Generalize for multiple federal agencies such as USDA and NSF, as well as international organizations such as Deutsche Bundesbank
- Bring in AI expertise from industry and academia: KAIST, LARC, Recognai, DLA, Primer AI, GESIS, AllenAI, etc.

Agency Benefits

- Understand more about agency's user community, to help outreach and get feedback, especially for researchers or commercial entities using agency data in novel ways.
- Help quantify the value and impact of data and research.
- Relevance w.r.t. the *Federal Data Strategy* and the *OPEN Government Data Act*: comprehensive data inventory, tools to help data users find trustworthy and relevant data.
- Explore a novel solution to dataset search.

Additional Information

Rich Context @ NYU Coleridge Initiative
coleridgeinitiative.org/richcontext

- **white paper**
- **upcoming book** (Jan 2020)
- **feedback/propose collaboration**

"Empty rhetoric over data sharing slows science"
Nature (2017-06-12)

"Experiences of the Deutsche Bundesbank"
Stefan Bender
CEMLA (2019-05-28)

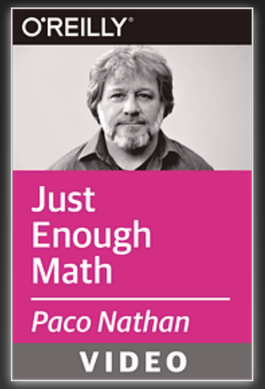
"Where's Waldo: Finding datasets in empirical research publications"
Julia Lane
AKBC (2019-05-22)

"Google data set search"
Ian Mulvany
ScholCommsBlog (2019-11-19)

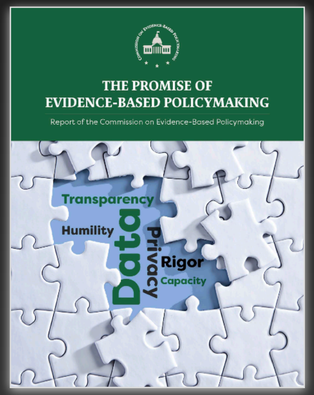
"Impact for social science researchers"
Ian Mulvany
FORCE11 (2019-11-17)

publications, interviews, conference summaries…

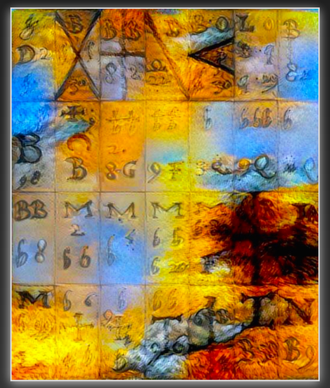
https://derwen.ai/paco
@pacoid



Just Enough Math



Rich Context



Hylbert-Speys

Rev conf

Themes + Confs
per Pacoid

derwen.ai