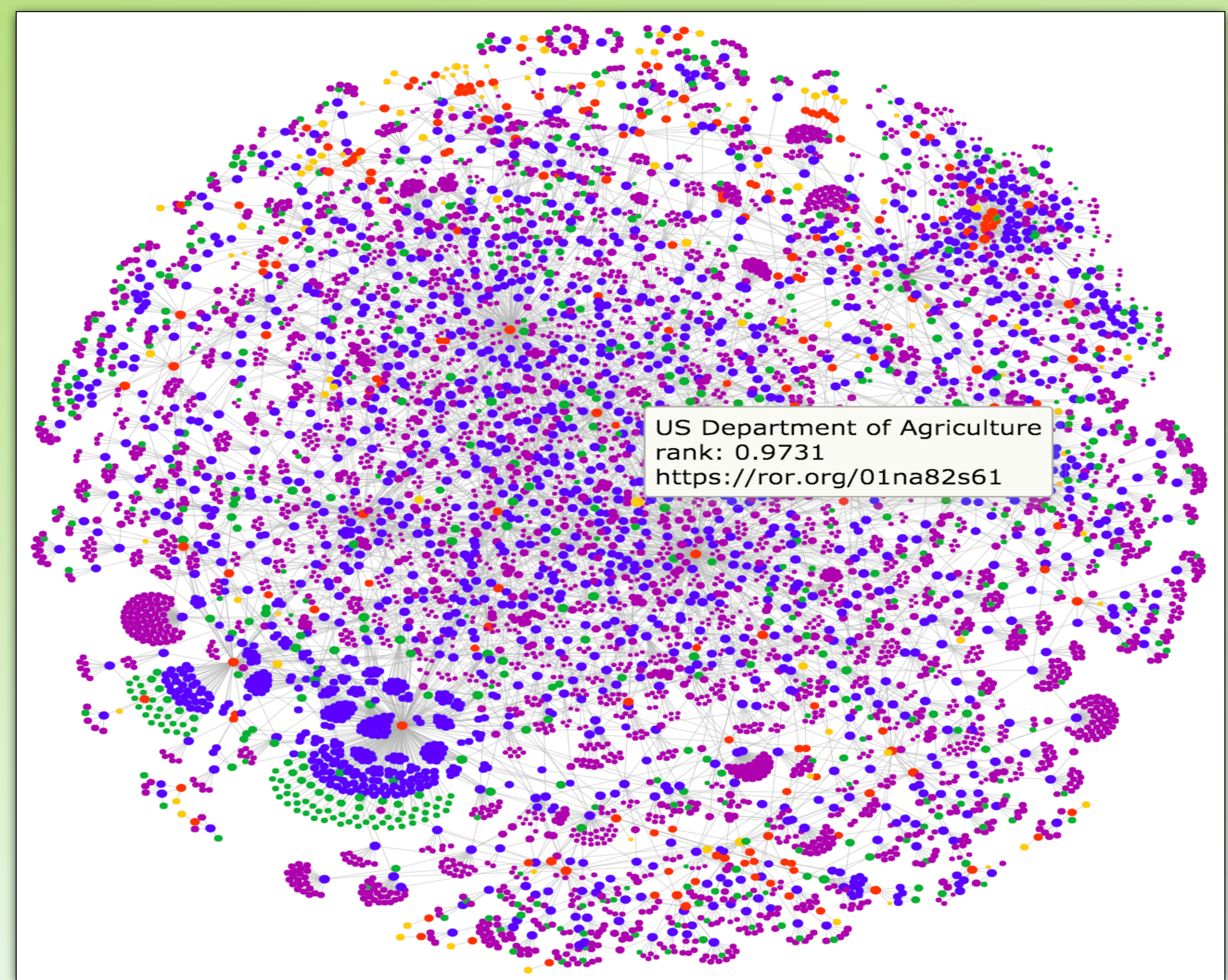
# Rich Context:

## Rich Search and Discovery for Scholarly Datasets

Julia Lane

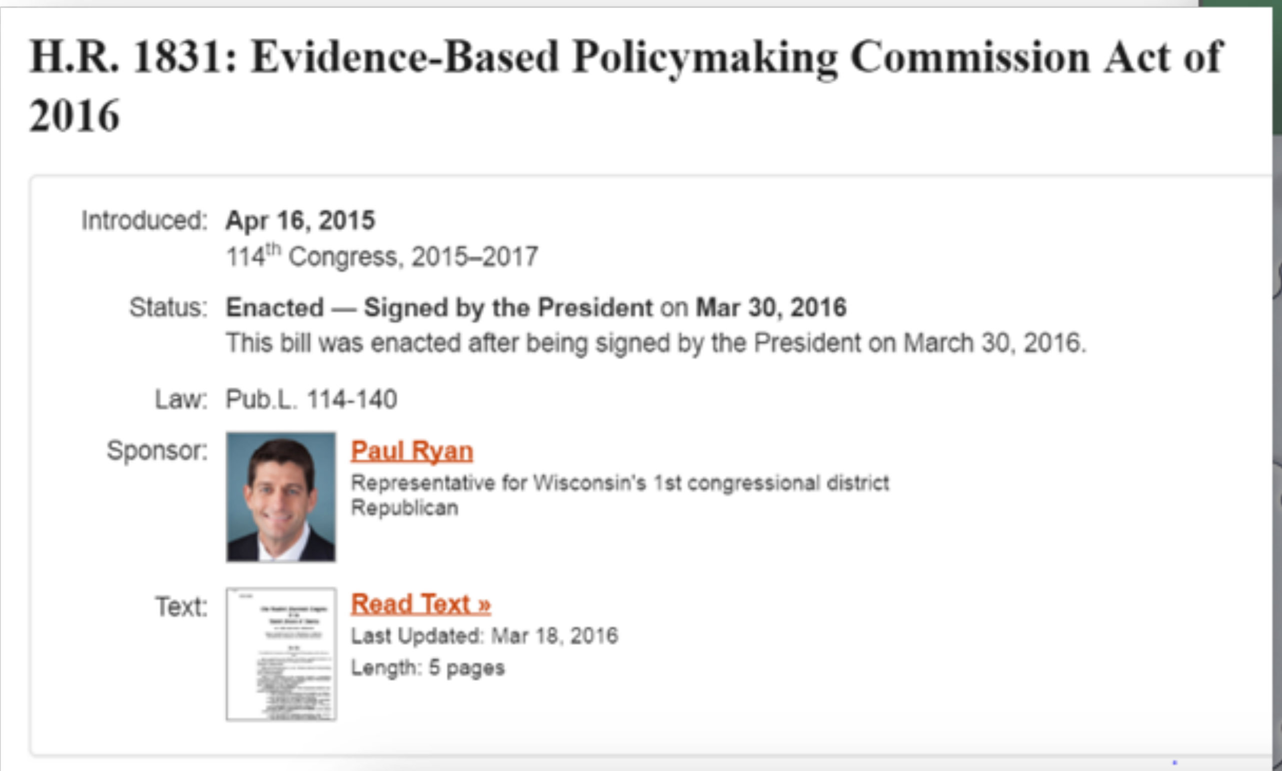
Paco Nathan

NYU Coleridge Initiative



US Department of Agriculture
rank: 0.9731
https://ror.org/01na82s61

# Context



## H.R. 1831: Evidence-Based Policymaking Commission Act of 2016

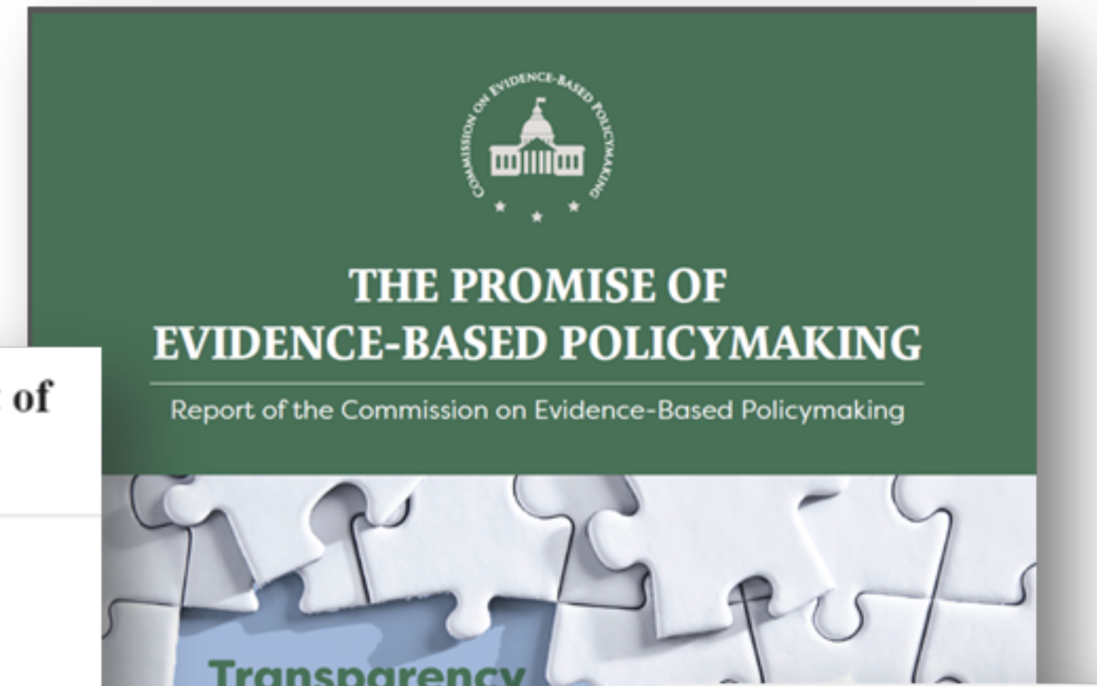
Introduced: **Apr 16, 2015**
114th Congress, 2015–2017

Status: **Enacted — Signed by the President** on **Mar 30, 2016**
This bill was enacted after being signed by the President on March 30, 2016.

Law: Pub.L. 114-140

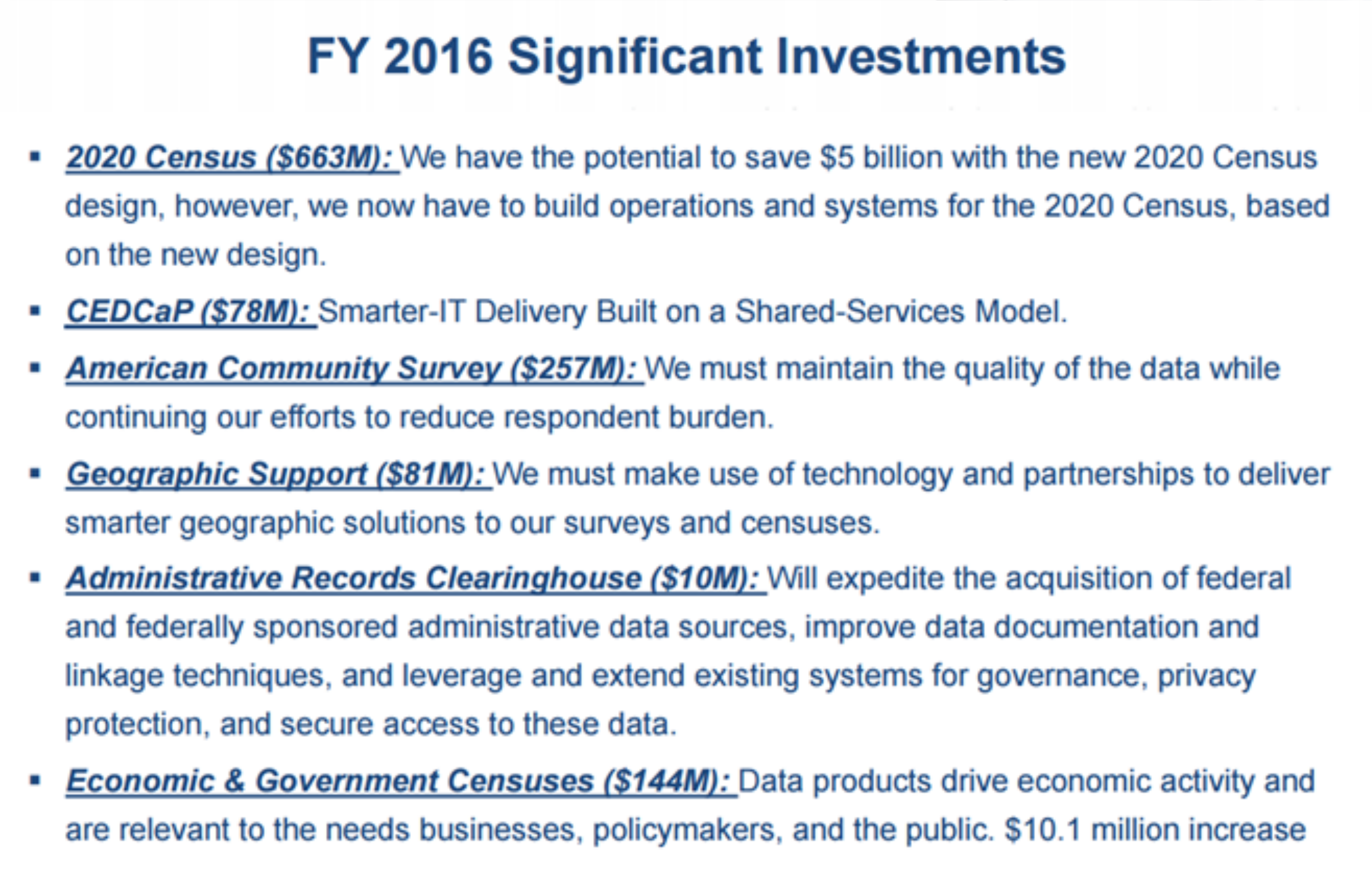
Sponsor: Paul Ryan
Representative for Wisconsin's 1st congressional district
Republican

Text: Read Text »
Last Updated: Mar 18, 2016
Length: 5 pages



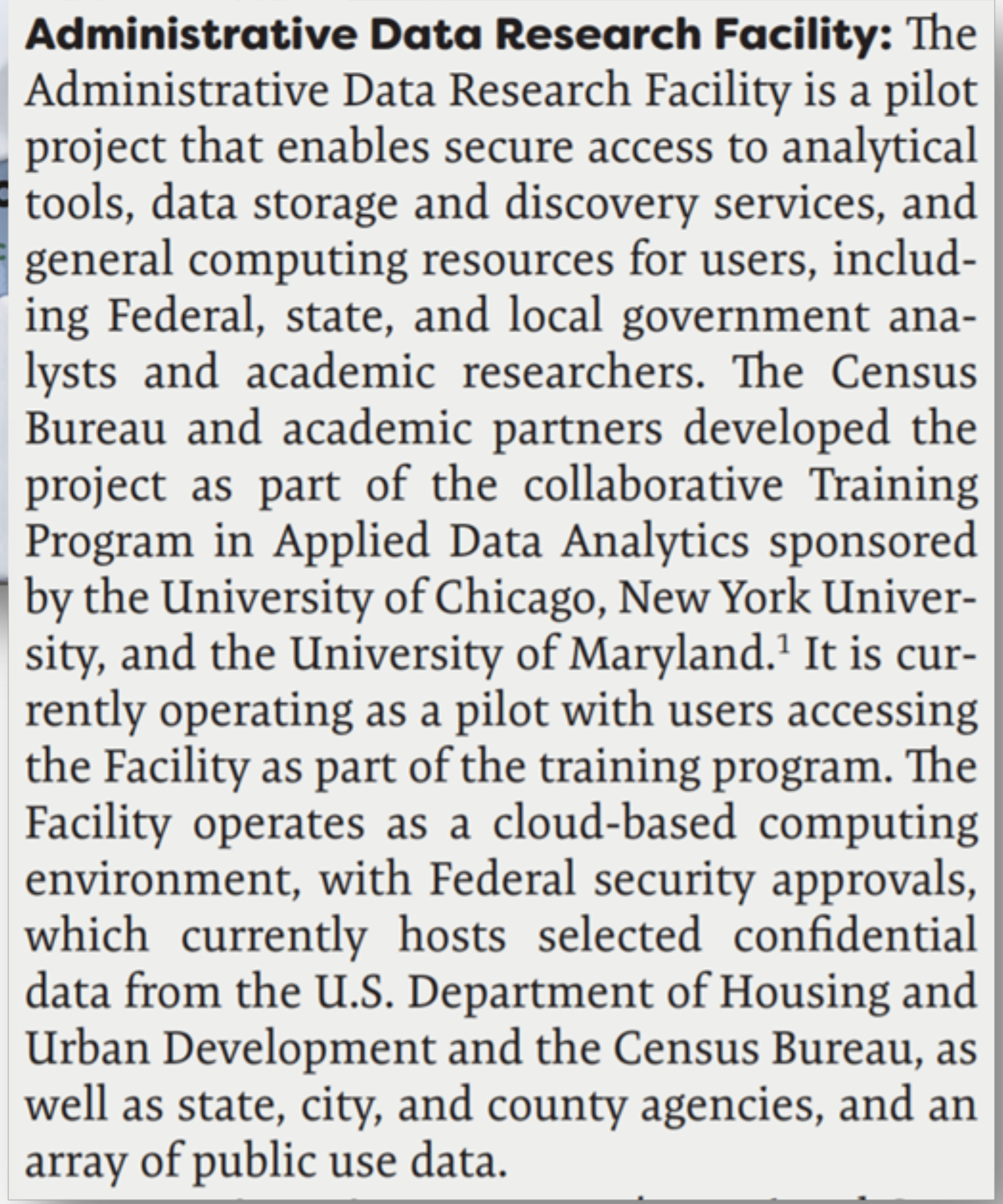
THE PROMISE OF
EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking



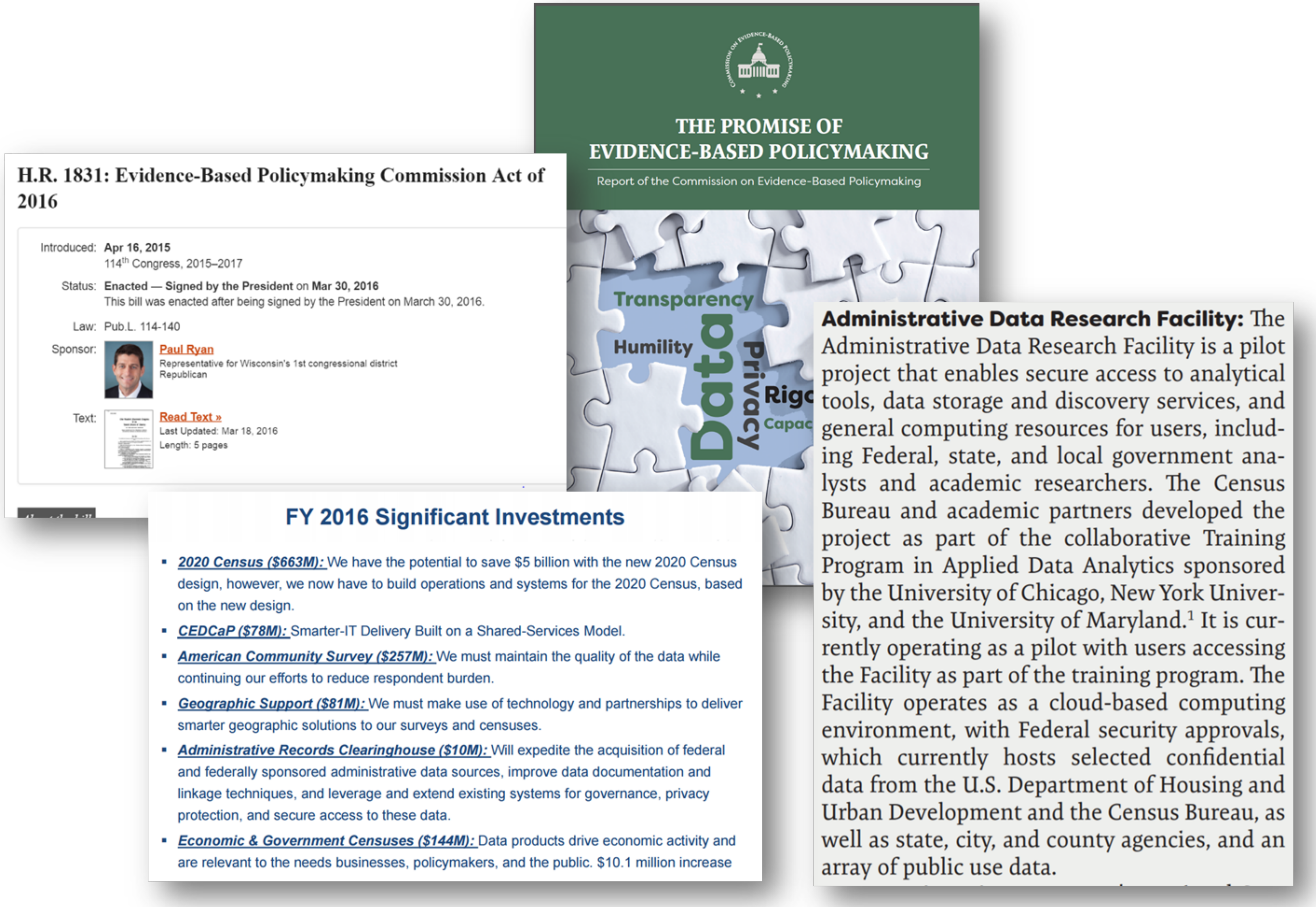
## FY 2016 Significant Investments

- **2020 Census ($663M):** We have the potential to save $5 billion with the new 2020 Census design, however, we now have to build operations and systems for the 2020 Census, based on the new design.
- **CEDCaP ($78M):** Smarter-IT Delivery Built on a Shared-Services Model.
- **American Community Survey ($257M):** We must maintain the quality of the data while continuing our efforts to reduce respondent burden.
- **Geographic Support ($81M):** We must make use of technology and partnerships to deliver smarter geographic solutions to our surveys and censuses.
- **Administrative Records Clearinghouse ($10M):** Will expedite the acquisition of federal and federally sponsored administrative data sources, improve data documentation and linkage techniques, and leverage and extend existing systems for governance, privacy protection, and secure access to these data.
- **Economic & Government Censuses ($144M):** Data products drive economic activity and are relevant to the needs businesses, policymakers, and the public. $10.1 million increase



**Administrative Data Research Facility:** The Administrative Data Research Facility is a pilot project that enables secure access to analytical tools, data storage and discovery services, and general computing resources for users, including Federal, state, and local government analysts and academic researchers. The Census Bureau and academic partners developed the project as part of the collaborative Training Program in Applied Data Analytics sponsored by the University of Chicago, New York University, and the University of Maryland.[1] It is currently operating as a pilot with users accessing the Facility as part of the training program. The Facility operates as a cloud-based computing environment, with Federal security approvals, which currently hosts selected confidential data from the U.S. Department of Housing and Urban Development and the Census Bureau, as well as state, city, and county agencies, and an array of public use data.

# Context



H.R. 1831: Evidence-Based Policymaking Commission Act of 2016

Introduced: Apr 16, 2015
114th Congress, 2015–2017

Status: Enacted — Signed by the President on Mar 30, 2016
This bill was enacted after being signed by the President on March 30, 2016.

Law: Pub.L. 114-140

Sponsor: Paul Ryan
Representative for Wisconsin's 1st congressional district
Republican

Text: Read Text »
Last Updated: Mar 18, 2016
Length: 5 pages

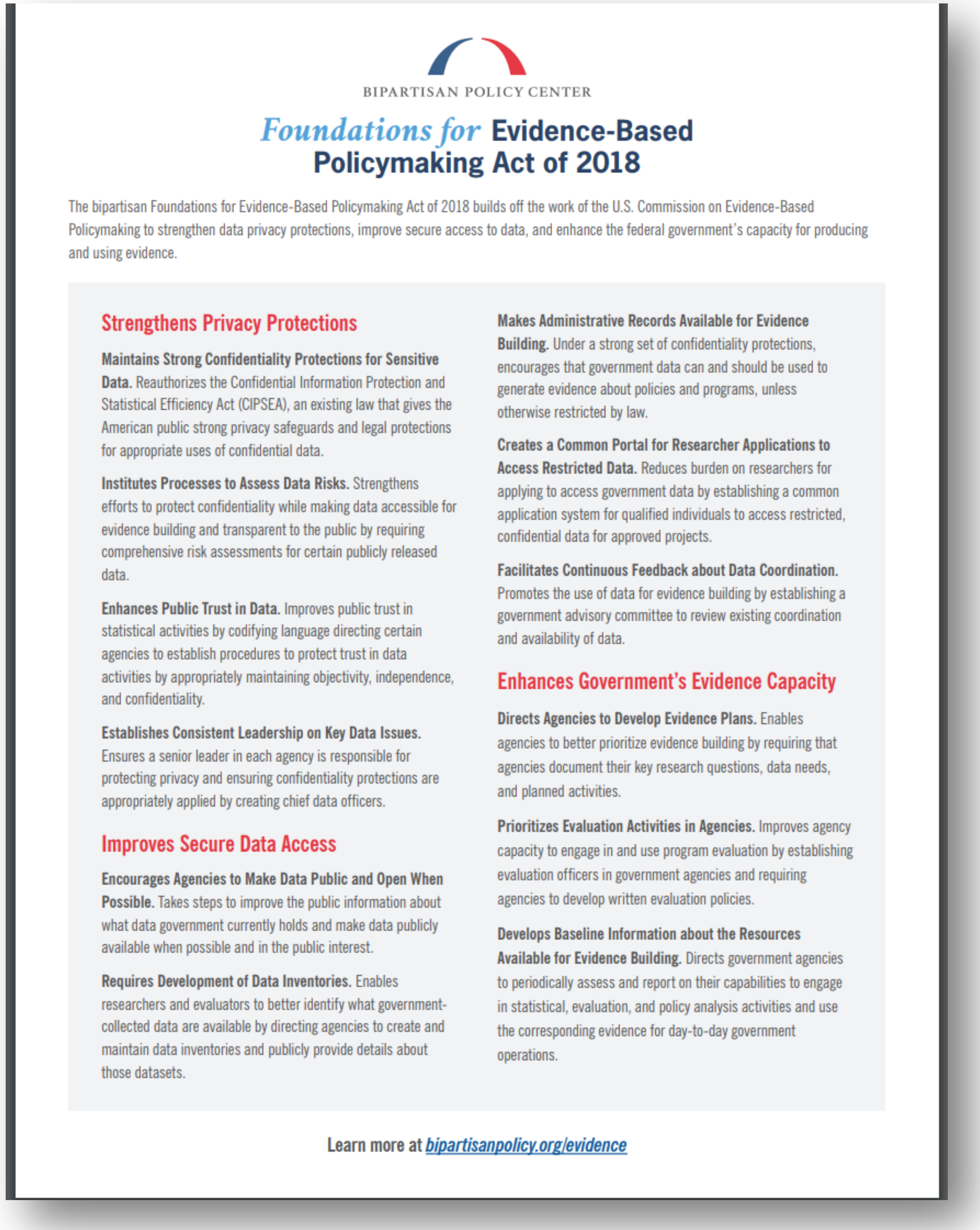
THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

**FY 2016 Significant Investments**

- 2020 Census ($663M): We have the potential to save $5 billion with the new 2020 Census design, however, we now have to build operations and systems for the 2020 Census, based on the new design.
- CEDCaP ($78M): Smarter-IT Delivery Built on a Shared-Services Model.
- American Community Survey ($257M): We must maintain the quality of the data while continuing our efforts to reduce respondent burden.
- Geographic Support ($81M): We must make use of technology and partnerships to deliver smarter geographic solutions to our surveys and censuses.
- Administrative Records Clearinghouse ($10M): Will expedite the acquisition of federal and federally sponsored administrative data sources, improve data documentation and linkage techniques, and leverage and extend existing systems for governance, privacy protection, and secure access to these data.
- Economic & Government Censuses ($144M): Data products drive economic activity and are relevant to the needs businesses, policymakers, and the public. $10.1 million increase

**Administrative Data Research Facility:** The Administrative Data Research Facility is a pilot project that enables secure access to analytical tools, data storage and discovery services, and general computing resources for users, including Federal, state, and local government analysts and academic researchers. The Census Bureau and academic partners developed the project as part of the collaborative Training Program in Applied Data Analytics sponsored by the University of Chicago, New York University, and the University of Maryland.[1] It is currently operating as a pilot with users accessing the Facility as part of the training program. The Facility operates as a cloud-based computing environment, with Federal security approvals, which currently hosts selected confidential data from the U.S. Department of Housing and Urban Development and the Census Bureau, as well as state, city, and county agencies, and an array of public use data.



BIPARTISAN POLICY CENTER

## *Foundations for* Evidence-Based Policymaking Act of 2018

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 builds off the work of the U.S. Commission on Evidence-Based Policymaking to strengthen data privacy protections, improve secure access to data, and enhance the federal government's capacity for producing and using evidence.

### Strengthens Privacy Protections

**Maintains Strong Confidentiality Protections for Sensitive Data.** Reauthorizes the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), an existing law that gives the American public strong privacy safeguards and legal protections for appropriate uses of confidential data.

**Institutes Processes to Assess Data Risks.** Strengthens efforts to protect confidentiality while making data accessible for evidence building and transparent to the public by requiring comprehensive risk assessments for certain publicly released data.

**Enhances Public Trust in Data.** Improves public trust in statistical activities by codifying language directing certain agencies to establish procedures to protect trust in data activities by appropriately maintaining objectivity, independence, and confidentiality.

**Establishes Consistent Leadership on Key Data Issues.** Ensures a senior leader in each agency is responsible for protecting privacy and ensuring confidentiality protections are appropriately applied by creating chief data officers.

### Improves Secure Data Access

**Encourages Agencies to Make Data Public and Open When Possible.** Takes steps to improve the public information about what data government currently holds and make data publicly available when possible and in the public interest.

**Requires Development of Data Inventories.** Enables researchers and evaluators to better identify what government-collected data are available by directing agencies to create and maintain data inventories and publicly provide details about those datasets.

**Makes Administrative Records Available for Evidence Building.** Under a strong set of confidentiality protections, encourages that government data can and should be used to generate evidence about policies and programs, unless otherwise restricted by law.

**Creates a Common Portal for Researcher Applications to Access Restricted Data.** Reduces burden on researchers for applying to access government data by establishing a common application system for qualified individuals to access restricted, confidential data for approved projects.

**Facilitates Continuous Feedback about Data Coordination.** Promotes the use of data for evidence building by establishing a government advisory committee to review existing coordination and availability of data.

### Enhances Government's Evidence Capacity

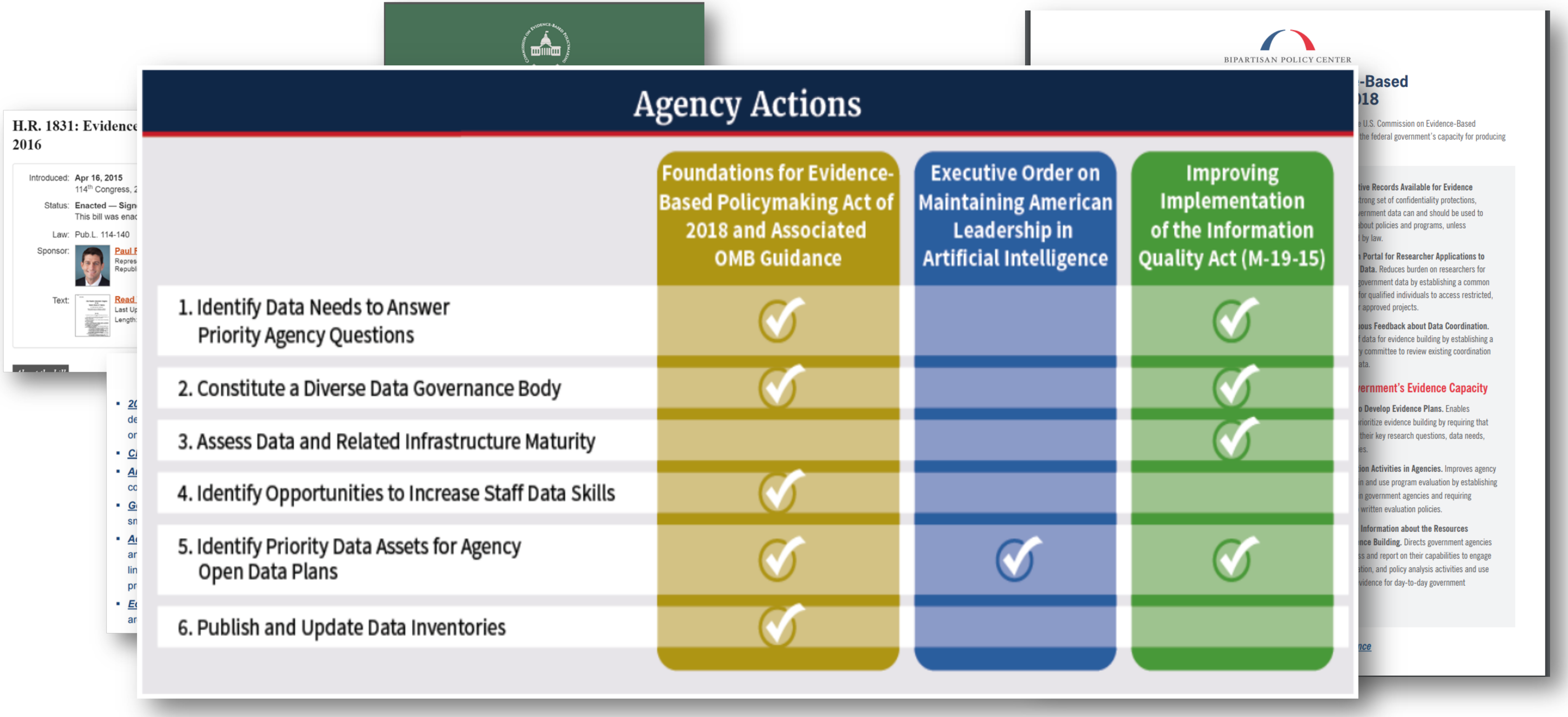
**Directs Agencies to Develop Evidence Plans.** Enables agencies to better prioritize evidence building by requiring that agencies document their key research questions, data needs, and planned activities.

**Prioritizes Evaluation Activities in Agencies.** Improves agency capacity to engage in and use program evaluation by establishing evaluation officers in government agencies and requiring agencies to develop written evaluation policies.

**Develops Baseline Information about the Resources Available for Evidence Building.** Directs government agencies to periodically assess and report on their capabilities to engage in statistical, evaluation, and policy analysis activities and use the corresponding evidence for day-to-day government operations.

Learn more at *bipartisanpolicy.org/evidence*

# Context



## Agency Actions

|  | Foundations for Evidence-Based Policymaking Act of 2018 and Associated OMB Guidance | Executive Order on Maintaining American Leadership in Artificial Intelligence | Improving Implementation of the Information Quality Act (M-19-15) |
|---|---|---|---|
| 1. Identify Data Needs to Answer Priority Agency Questions | ✓ |  | ✓ |
| 2. Constitute a Diverse Data Governance Body | ✓ |  | ✓ |
| 3. Assess Data and Related Infrastructure Maturity |  |  | ✓ |
| 4. Identify Opportunities to Increase Staff Data Skills | ✓ |  |  |
| 5. Identify Priority Data Assets for Agency Open Data Plans | ✓ | ✓ | ✓ |
| 6. Publish and Update Data Inventories | ✓ |  |  |

H.R. 1831: Evidence
2016

Introduced: Apr 16, 2015
114th Congress, 2

Status: Enacted — Sign
This bill was enac

Law: Pub.L. 114-140

Sponsor: Paul R
Repres
Republ

Text: Read
Last Up
Length:

BIPARTISAN POLICY CENTER

-Based
)18

e U.S. Commission on Evidence-Based
the federal government's capacity for producing

tive Records Available for Evidence
trong set of confidentiality protections,
ernment data can and should be used to
bout policies and programs, unless
l by law.

n Portal for Researcher Applications to
Data. Reduces burden on researchers for
government data by establishing a common
for qualified individuals to access restricted,
r approved projects.

uous Feedback about Data Coordination.
f data for evidence building by establishing a
y committee to review existing coordination
ata.

ernment's Evidence Capacity

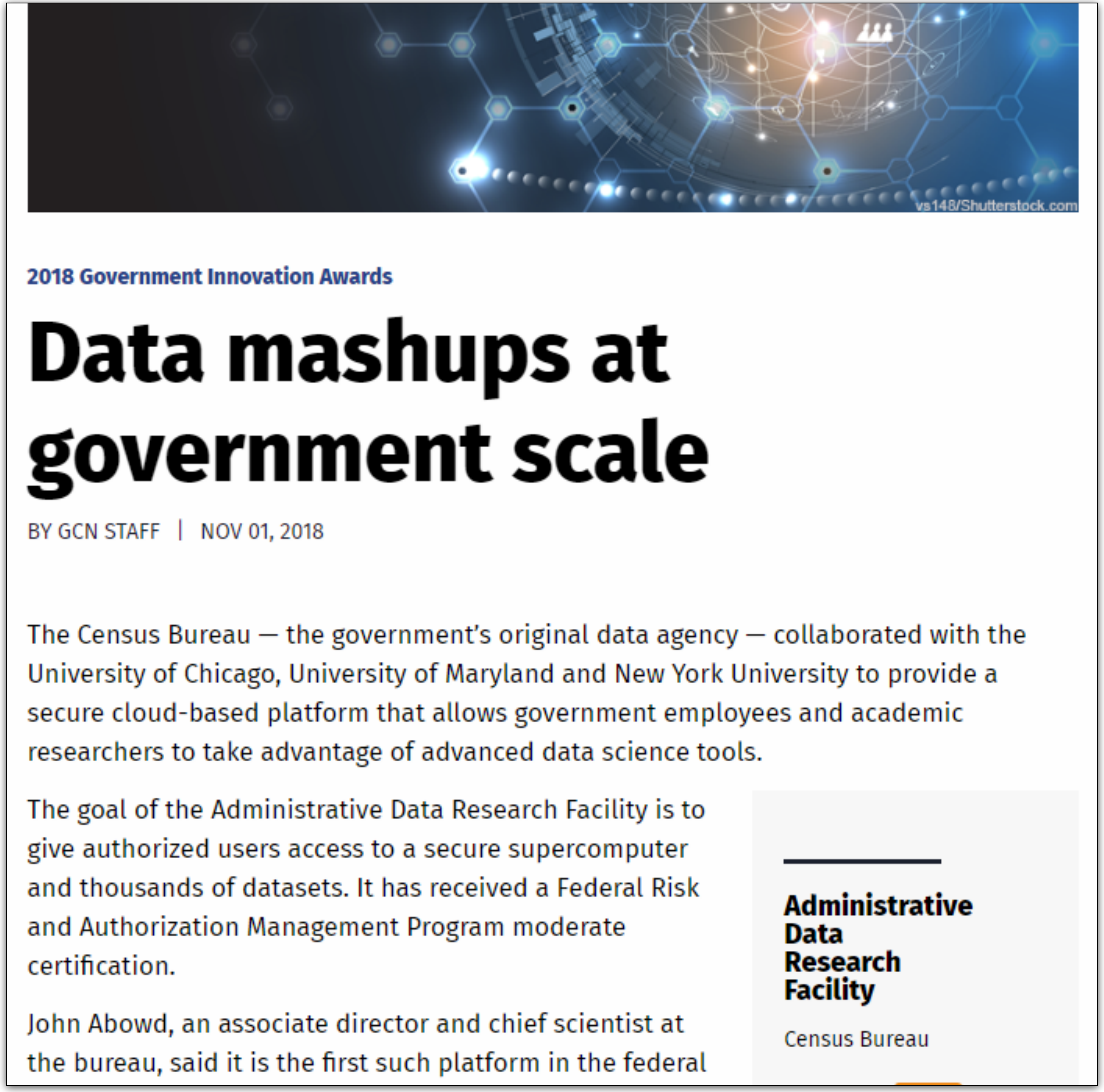
o Develop Evidence Plans. Enables
rioritize evidence building by requiring that
their key research questions, data needs,
es.

ion Activities in Agencies. Improves agency
n and use program evaluation by establishing
n government agencies and requiring
written evaluation policies.

Information about the Resources
nce Building. Directs government agencies
ss and report on their capabilities to engage
tion, and policy analysis activities and use
vidence for day-to-day government

# Evidence-Based Policymaking

- **Foundations for Evidence-based Policymaking Act** (2018)
- **Information Quality Act** (2001)
- **US Federal Data Strategy** (2019)
- **Year-1 Action Plan** (2019)



vs148/Shutterstock.com

2018 Government Innovation Awards

# Data mashups at government scale

BY GCN STAFF | NOV 01, 2018

The Census Bureau — the government's original data agency — collaborated with the University of Chicago, University of Maryland and New York University to provide a secure cloud-based platform that allows government employees and academic researchers to take advantage of advanced data science tools.

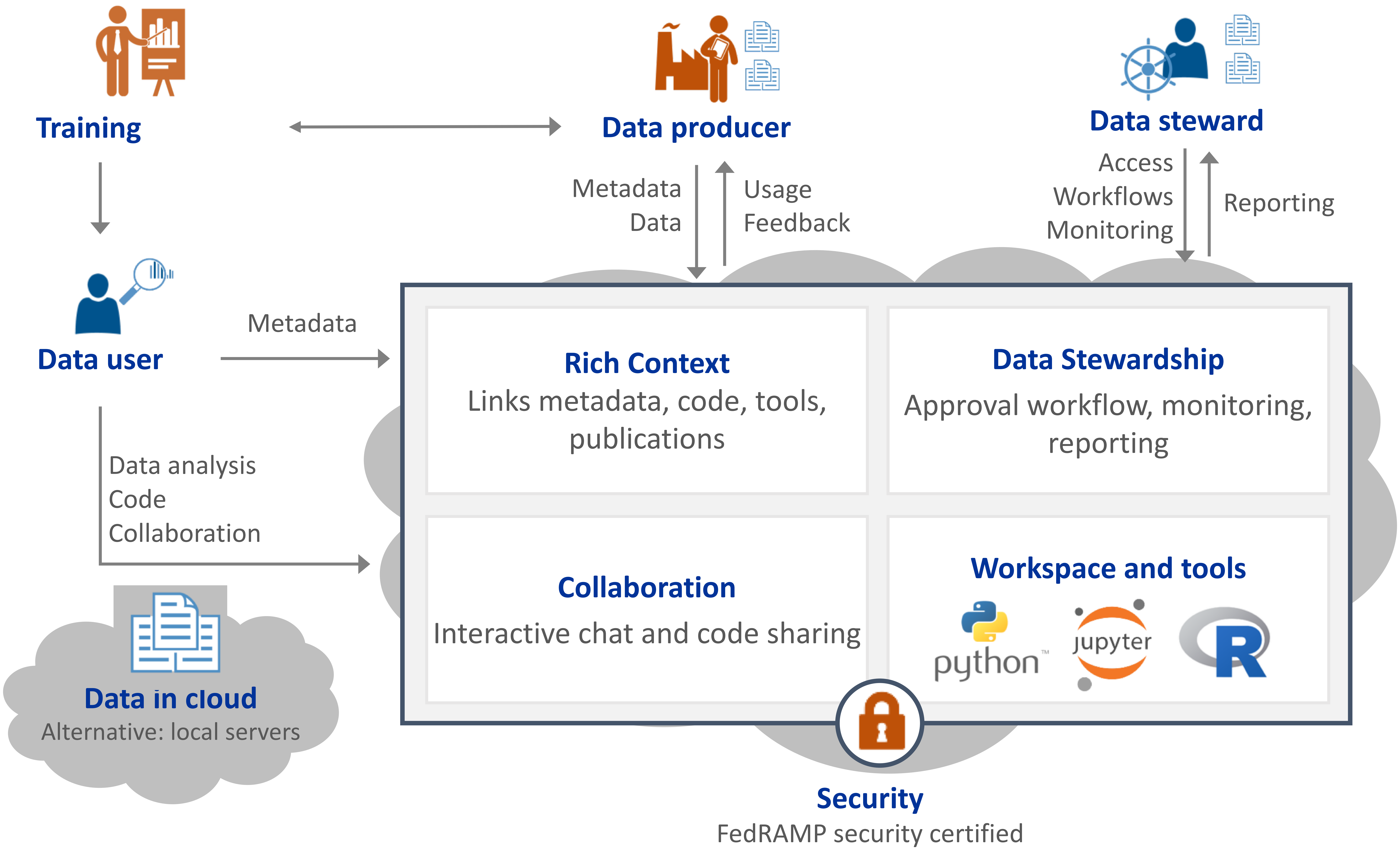
The goal of the Administrative Data Research Facility is to give authorized users access to a secure supercomputer and thousands of datasets. It has received a Federal Risk and Authorization Management Program moderate certification.

**Administrative Data Research Facility**

Census Bureau

John Abowd, an associate director and chief scientist at the bureau, said it is the first such platform in the federal

See article link



**Training**

**Data producer**

Metadata
Data

Usage
Feedback

**Data steward**

Access
Workflows
Monitoring

Reporting

**Data user**

Metadata

Data analysis
Code
Collaboration

**Data in cloud**
Alternative: local servers

**Rich Context**
Links metadata, code, tools, publications

**Data Stewardship**
Approval workflow, monitoring, reporting

**Collaboration**
Interactive chat and code sharing

**Workspace and tools**

**Security**
FedRAMP security certified

# 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

# Need: 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.

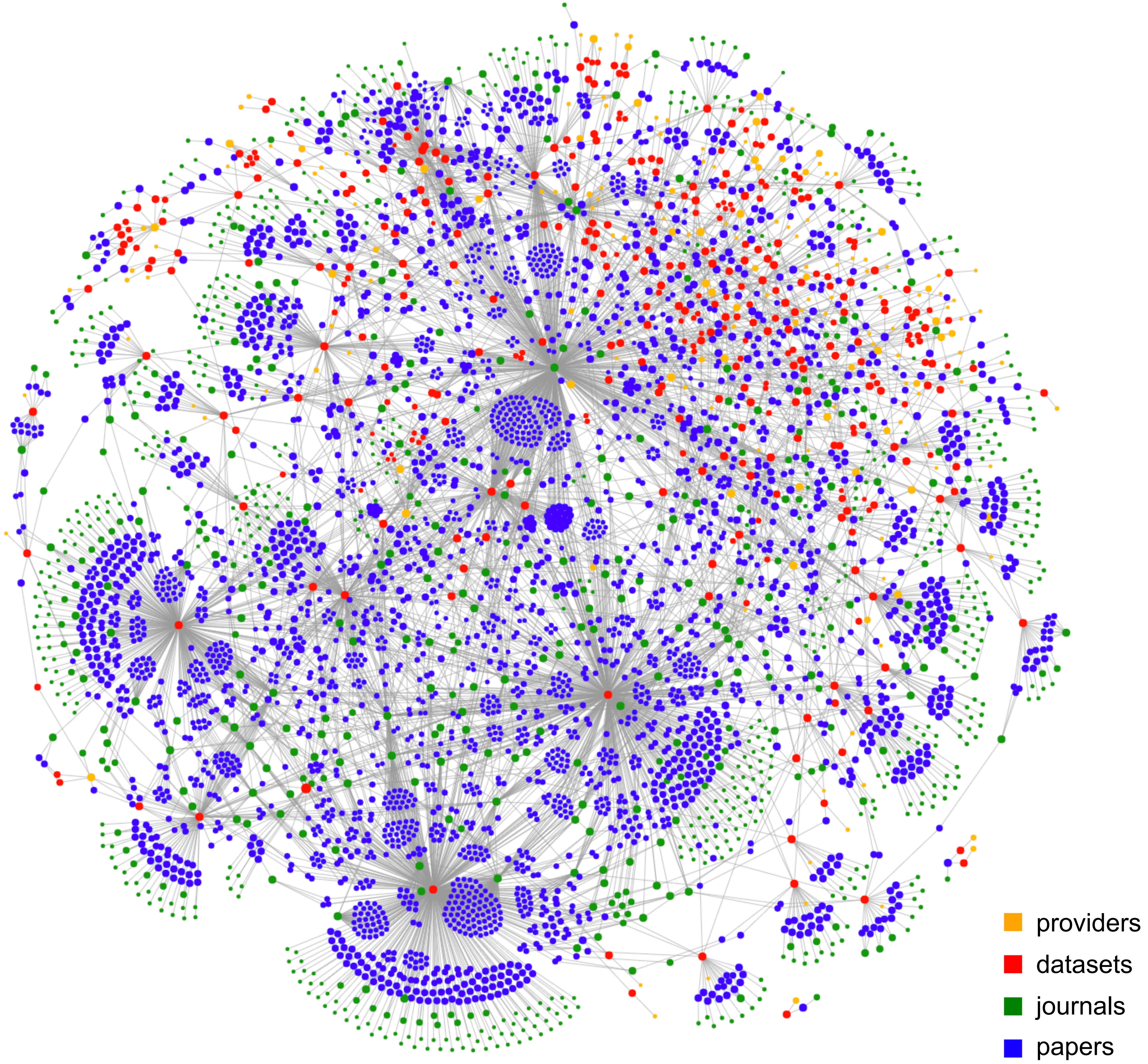
# Rich Context

- Focus on *socioeconomic impact*

- Funded by Schmidt Futures, Sloan, Overdeck

- Partnering with Bundesbank, USDA, etc.

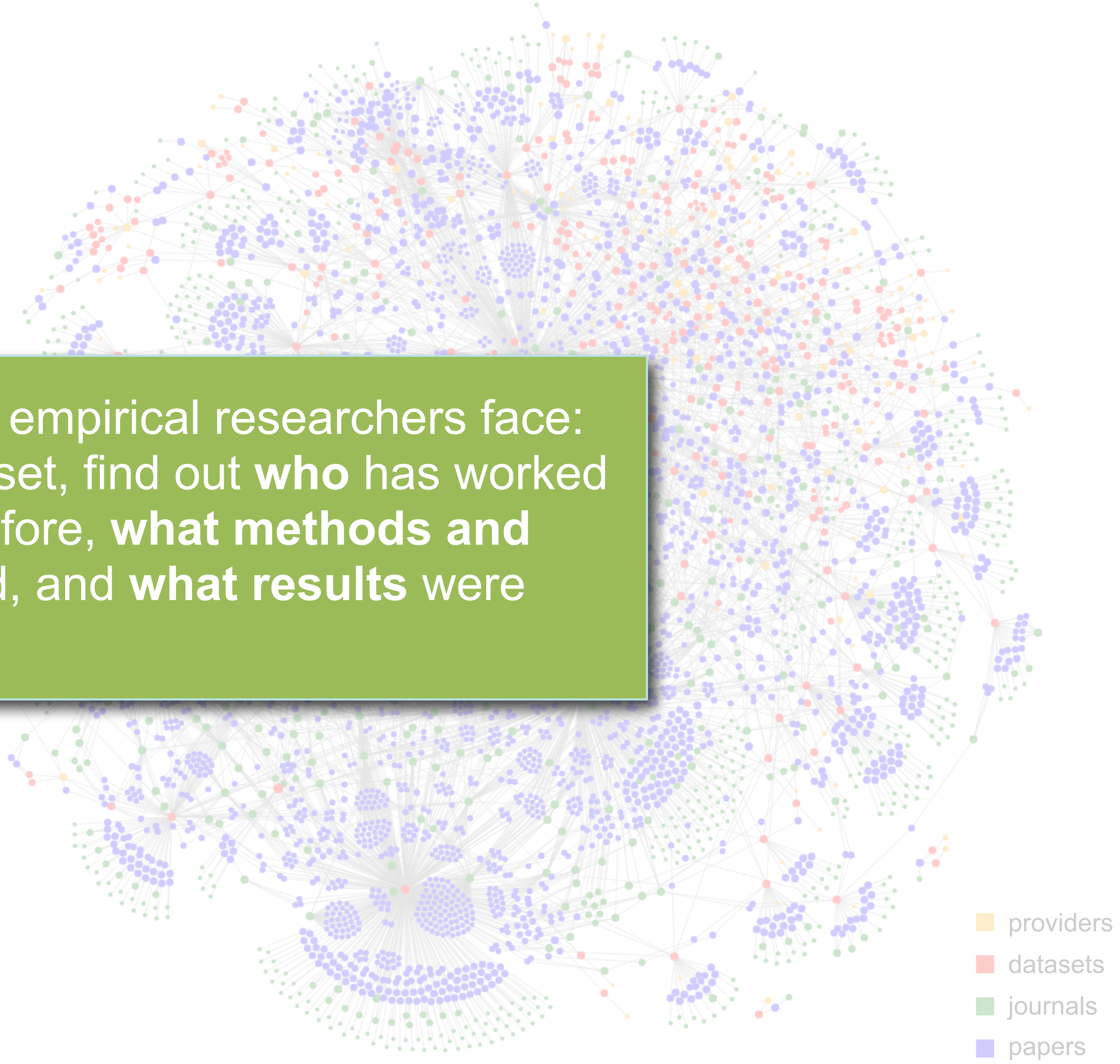
- Collaboration with SAGE Pub, RePEc, ResearchGate, Digital Science, etc.



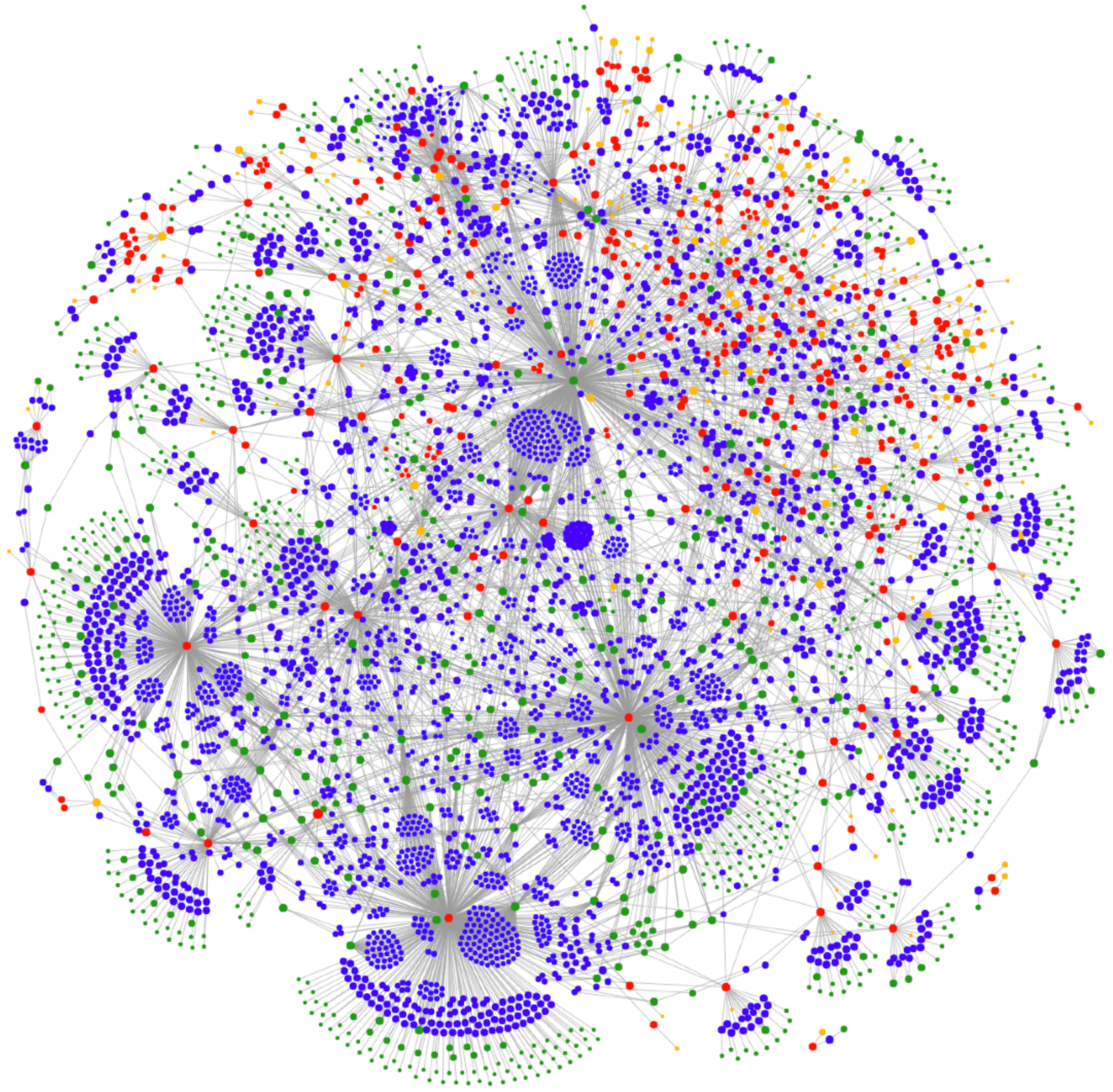
# Rich Context

- Focus on *socioeconomic impact*
- Funded by Schmidt Futures, Sloan, Overdec
- Partnering with USDA, etc.
- Collaboration w Pub, RePEc, R 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.

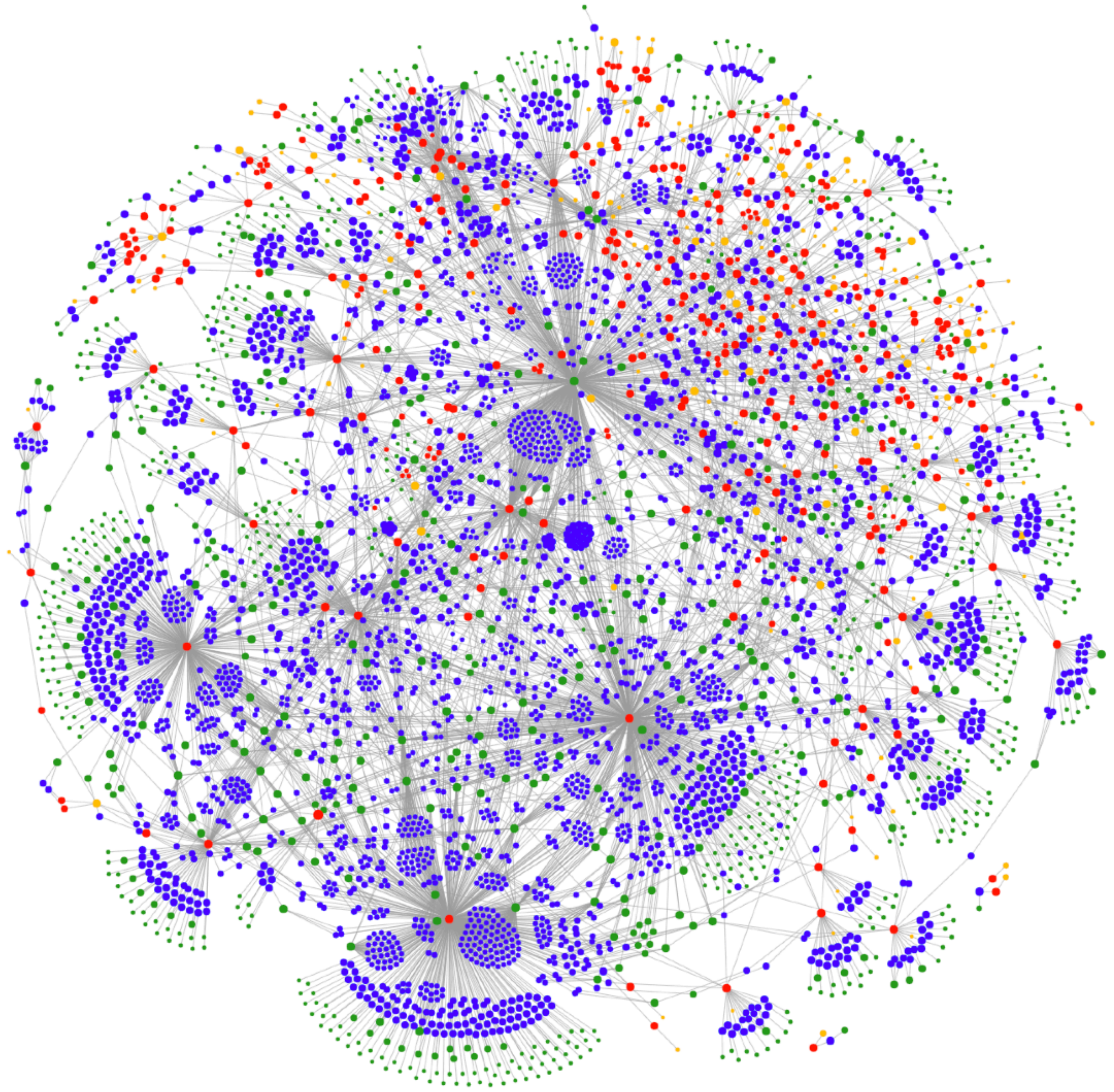


# Knowledge Graph – why?



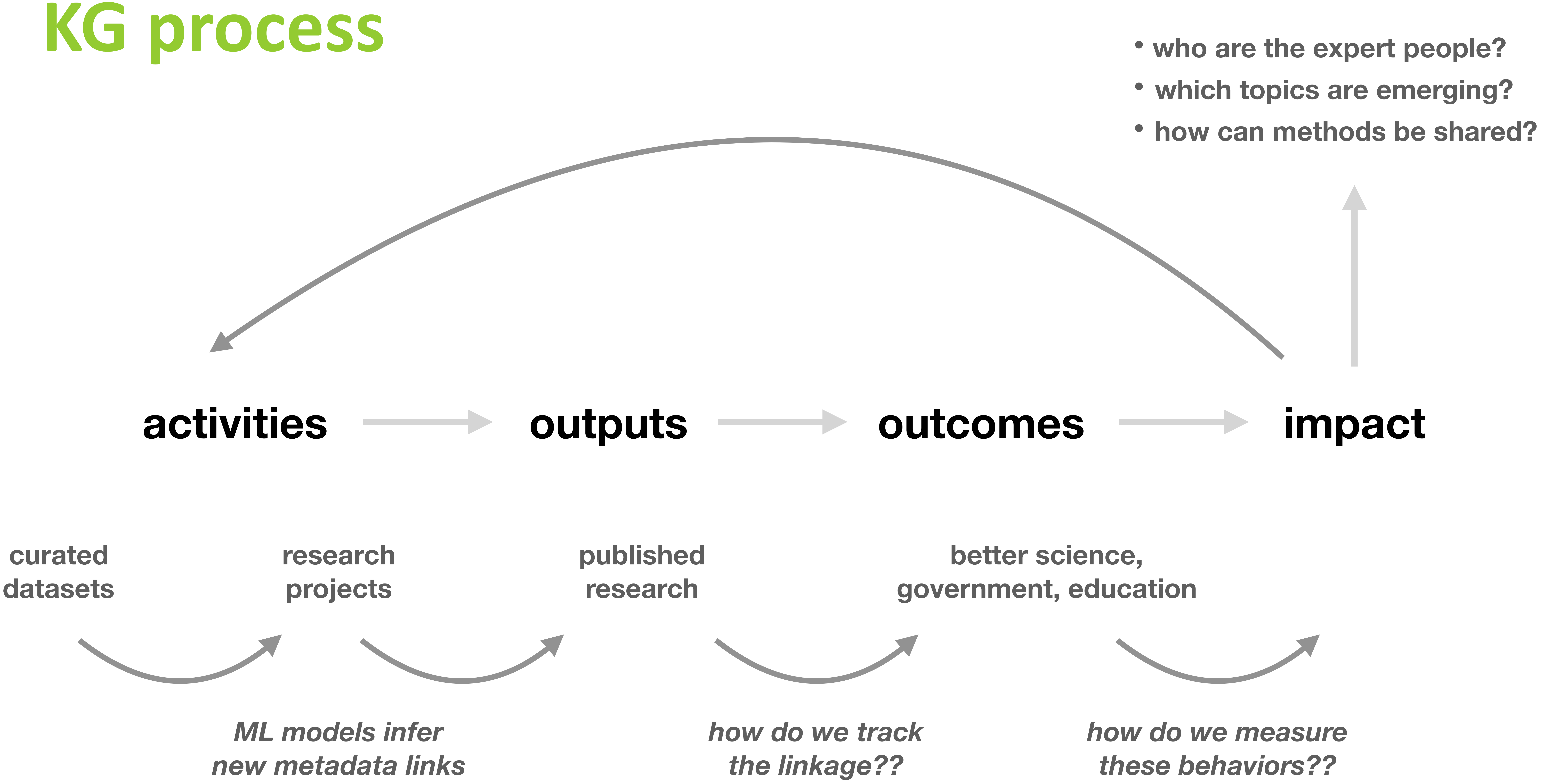
- Allow flexibility for metadata representation
- Measure metadata quality
- Prepare features for ML models
- Build recommenders for *experts*, *topics*, *tools*, etc.
- Engage the public with automated data inventories
- Recommend configurations to new analysts
- Identify which datasets get used with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

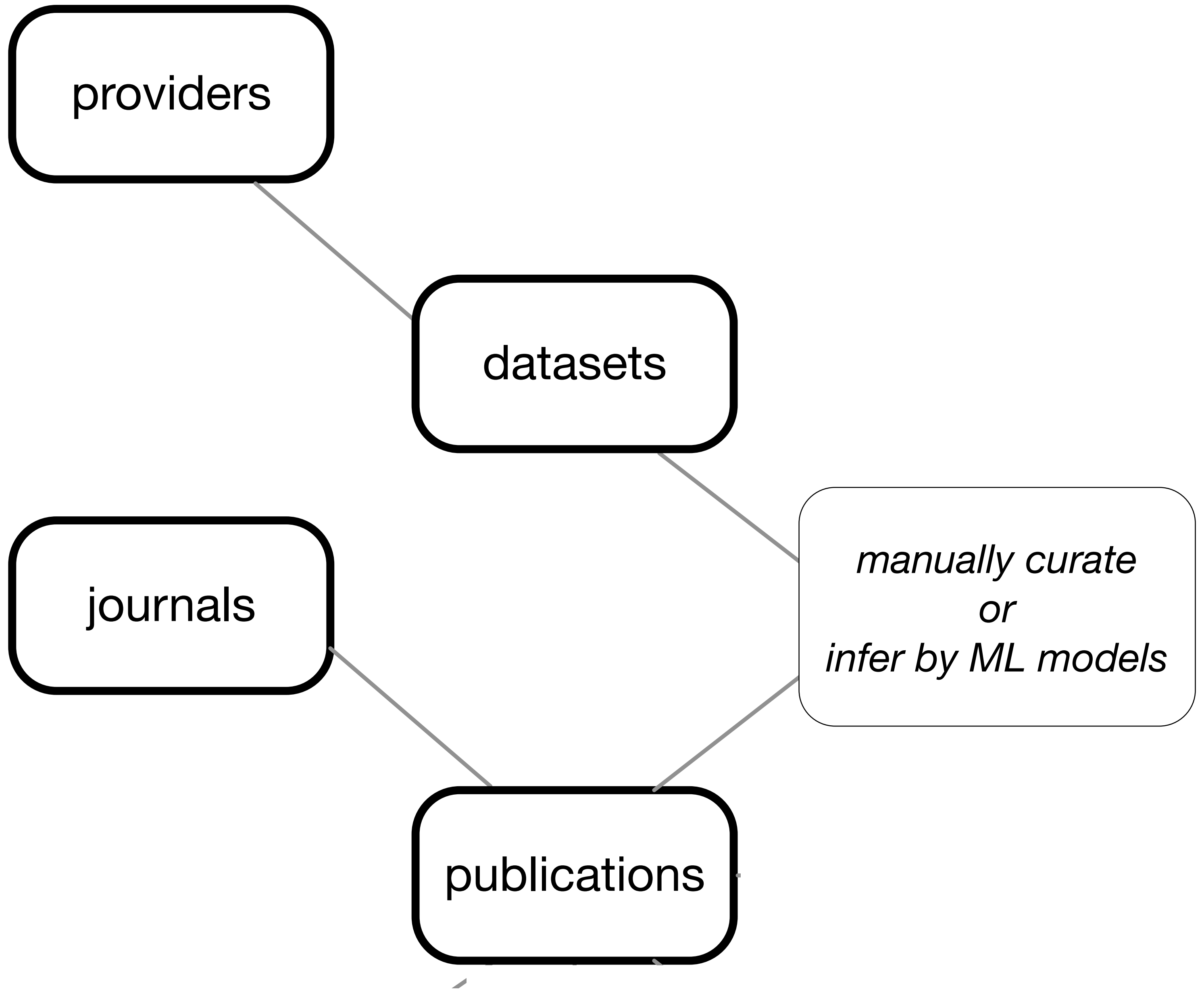
better science, government, education

*ML models infer new metadata links*

*how do we track the linkage??*

*how do we measure these behaviors??*

# KG construction and representation



providers

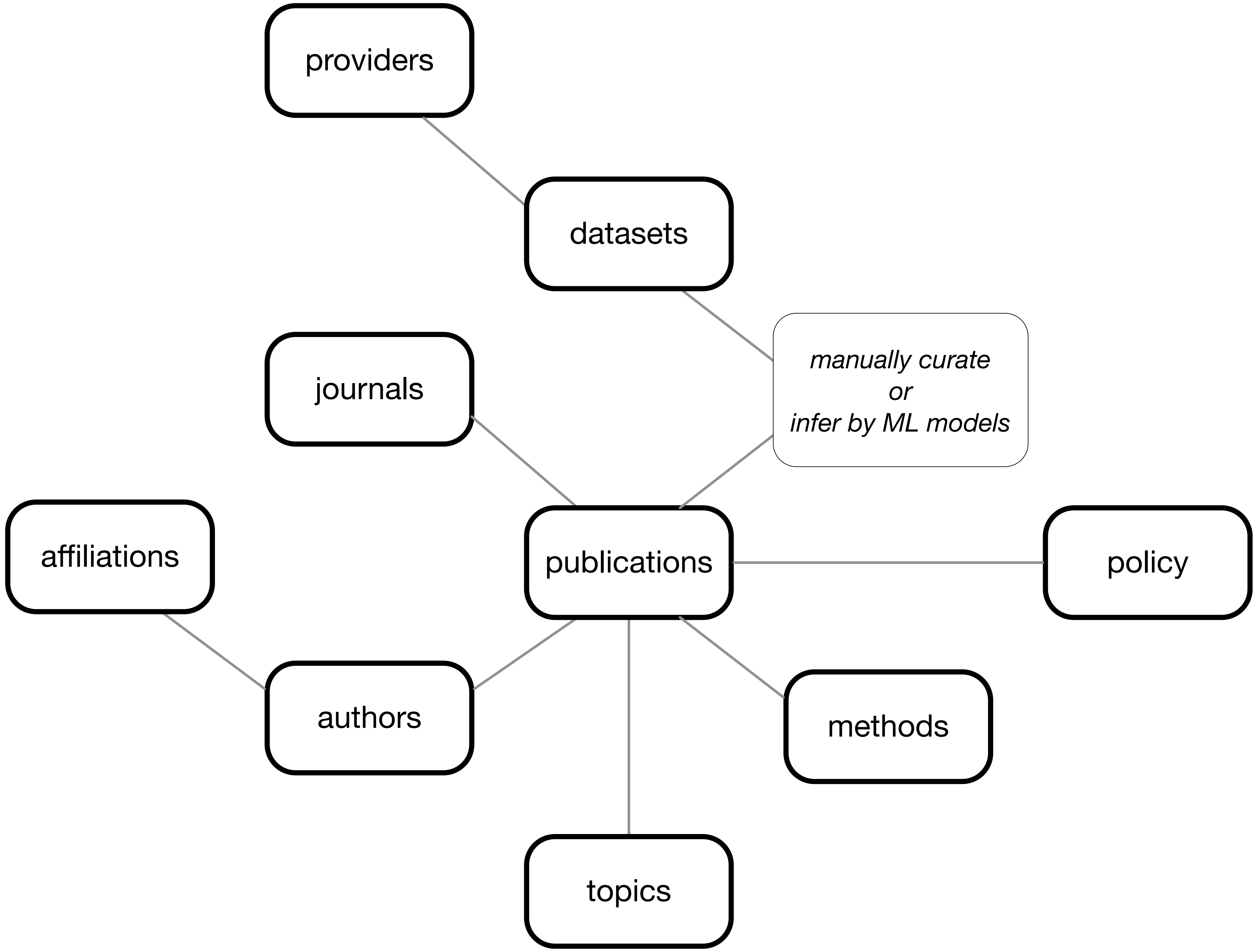
datasets

journals

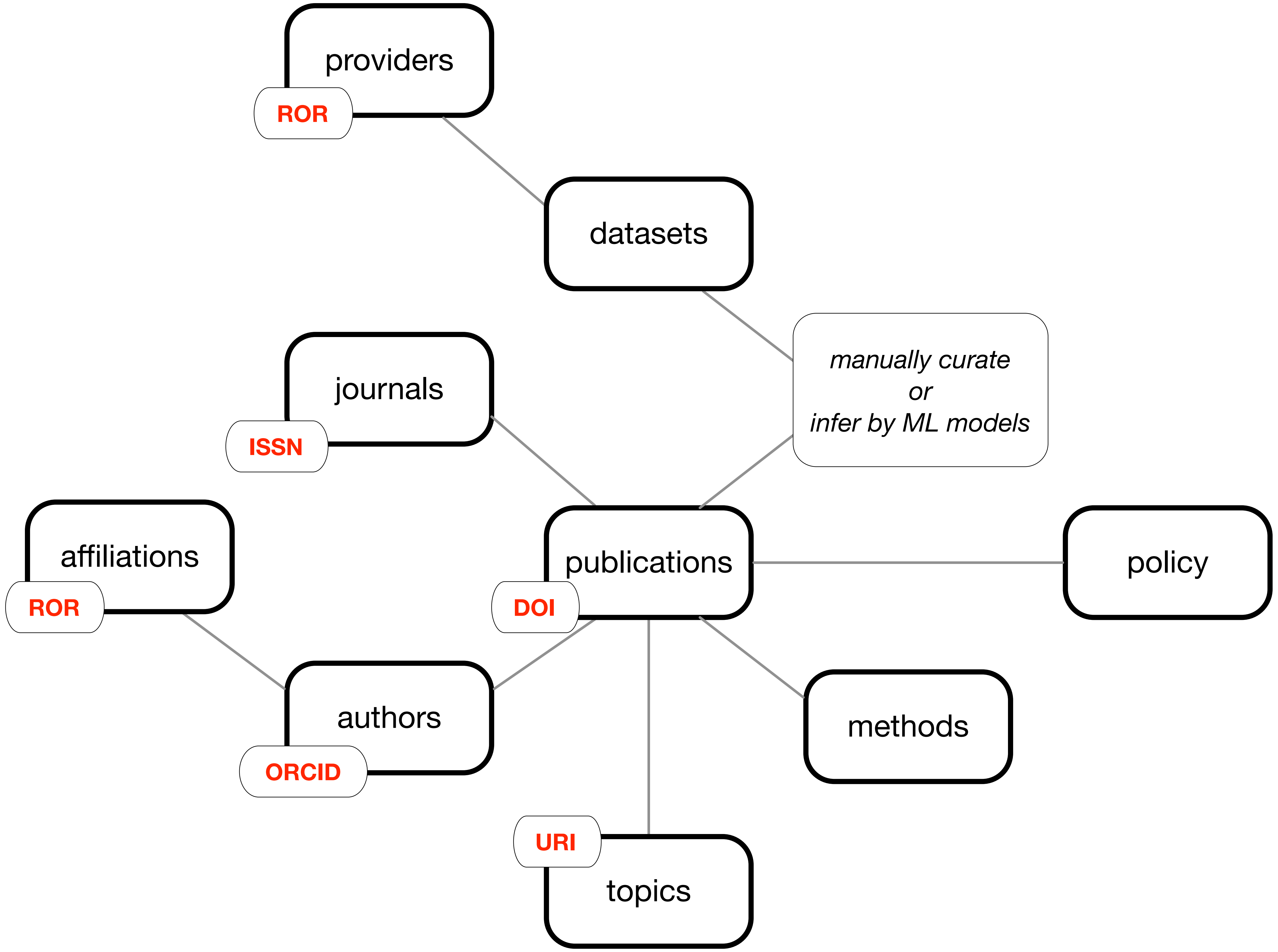
publications

*manually curate*
*or*
*infer by ML models*

# KG construction and representation

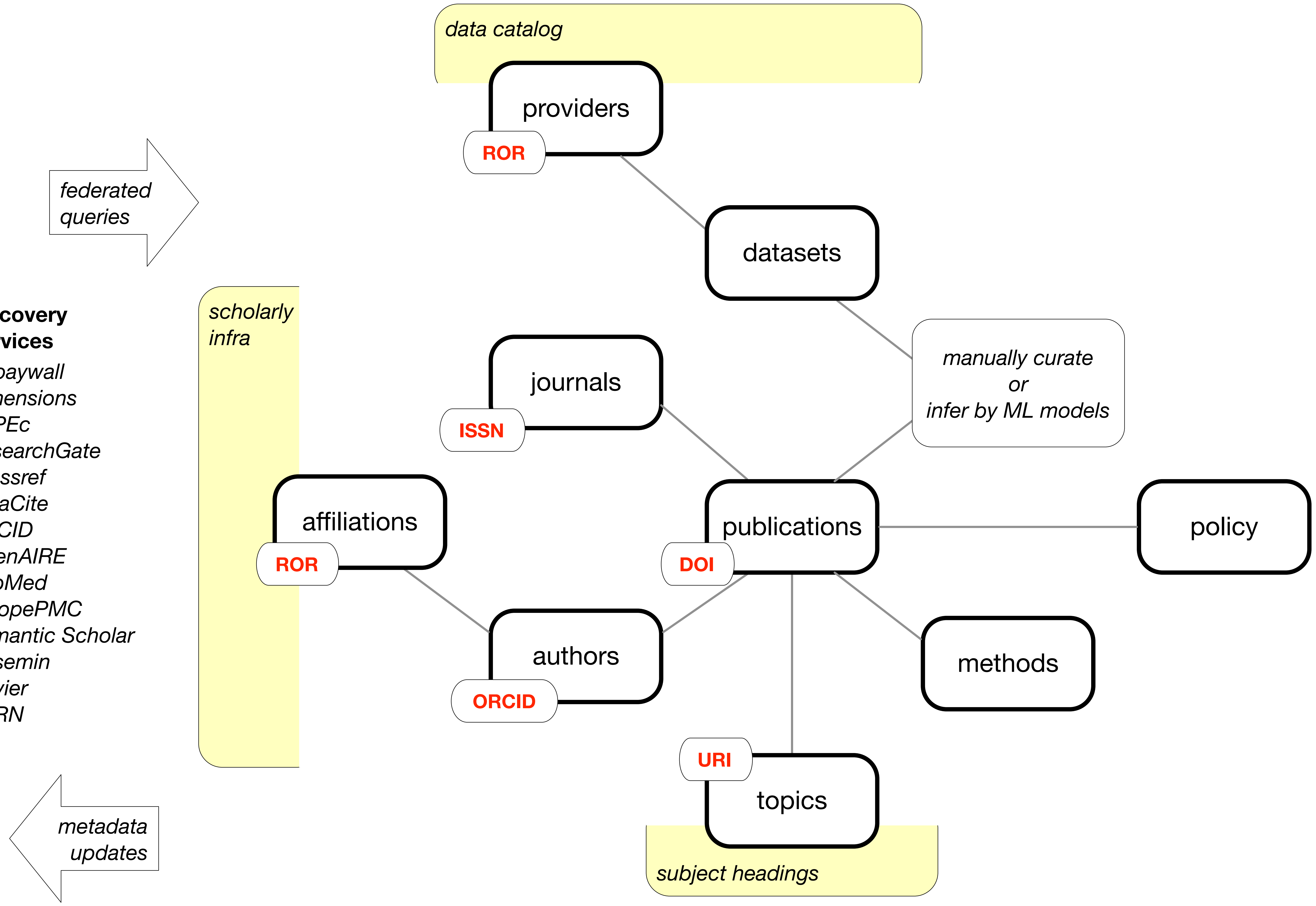


# KG construction and representation



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

# KG construction and representation



**Discovery Services**

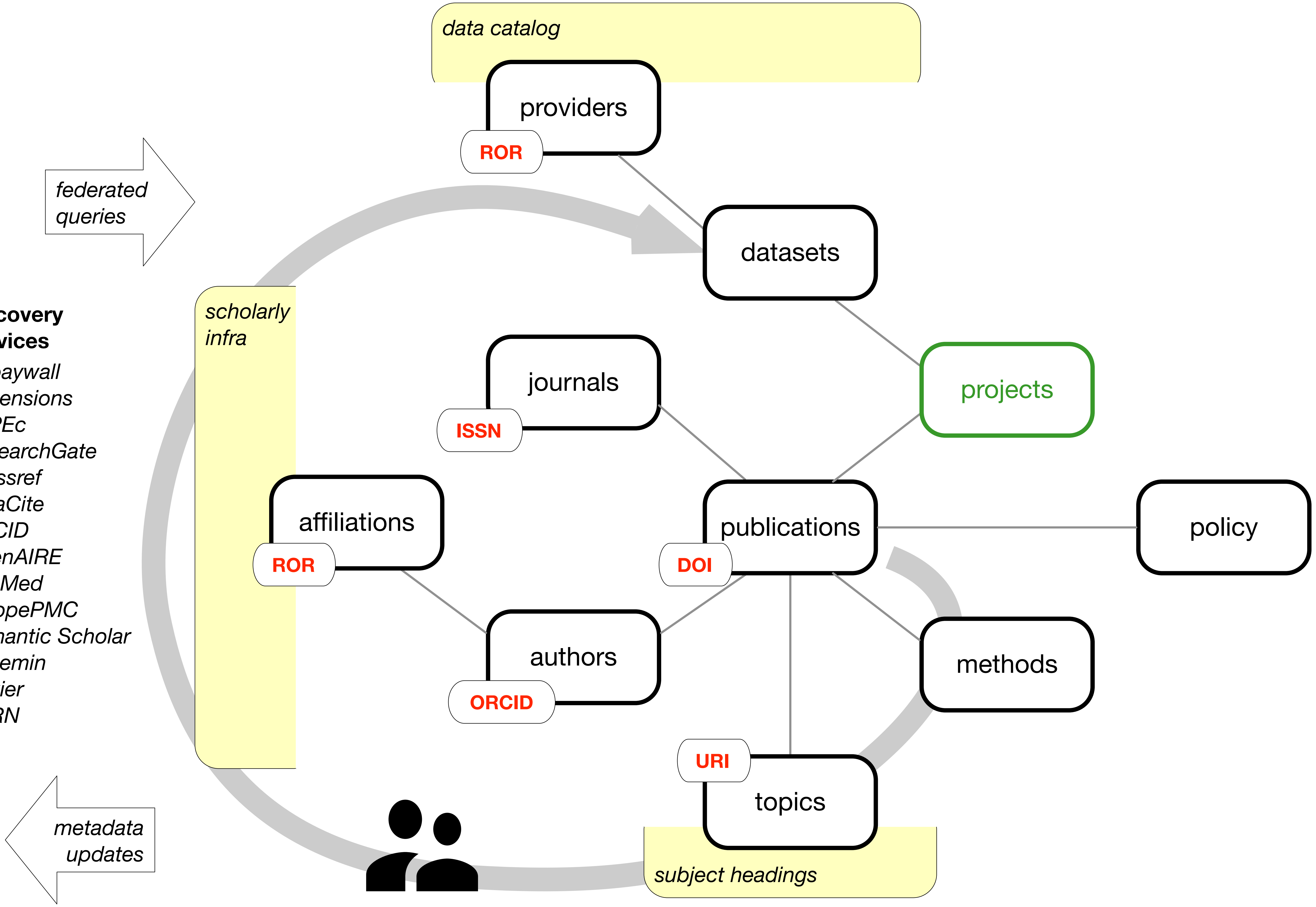
*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# KG construction and representation

**Discovery Services**

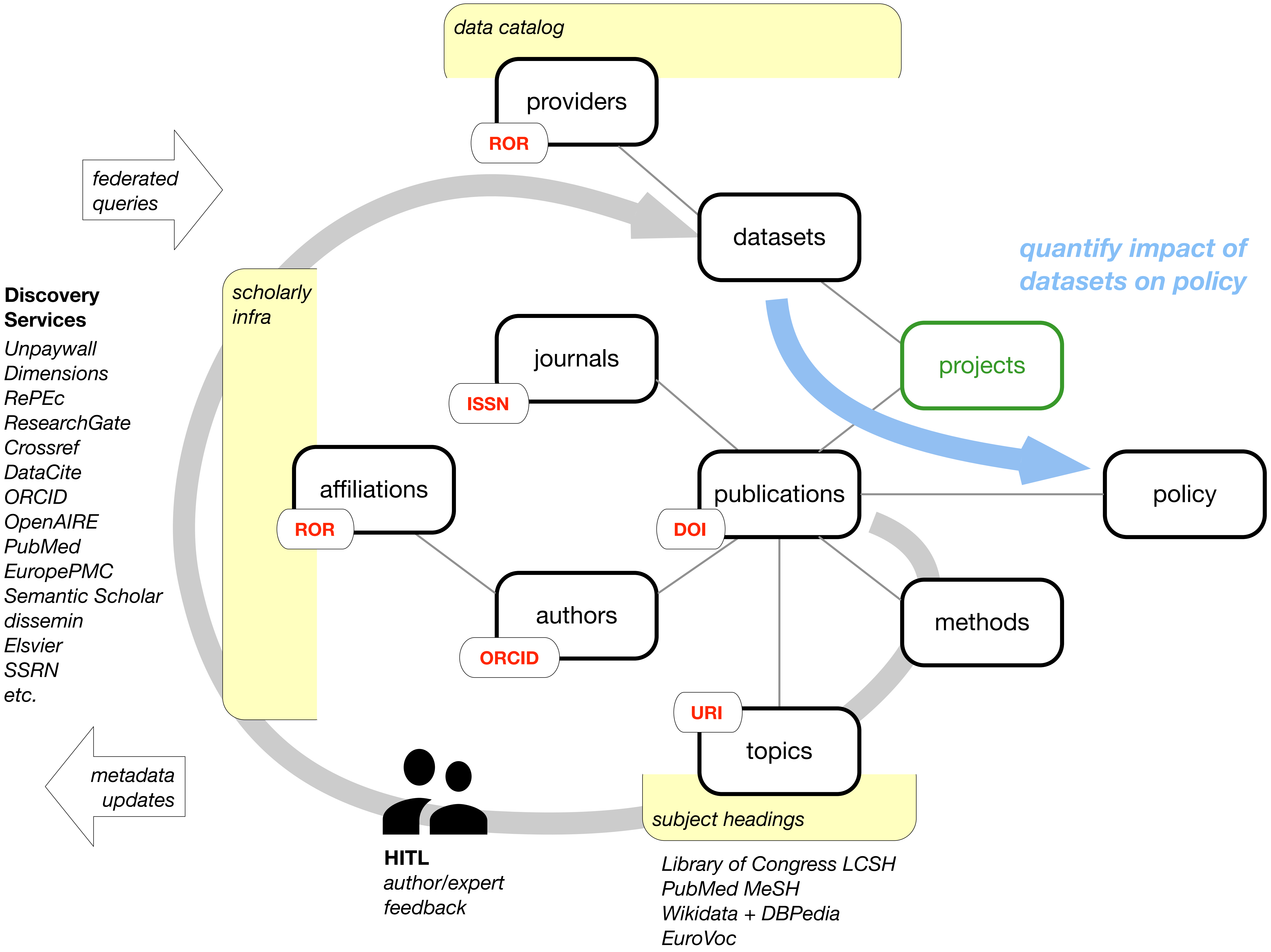
*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*



**HITL**
*author/expert feedback*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# KG construction and representation



*data catalog*

providers

ROR

*federated queries*

datasets

*quantify impact of datasets on policy*

**Discovery Services**

*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*

*scholarly infra*

journals

ISSN

projects

affiliations

ROR

publications

DOI

policy

authors

ORCID

methods

URI

topics

*metadata updates*

*subject headings*

**HITL**
*author/expert feedback*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# 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**
- **adrf-onto** – controlled vocabulary for ADRF and Rich Context using OWL, SKOS, DCAT, PAV, CITO, FaBiO, etc.
  **github.com/Coleridge-Initiative/adrf-onto**
- **RCLC** – ML leaderboard competition
  **github.com/Coleridge-Initiative/rclc**

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:

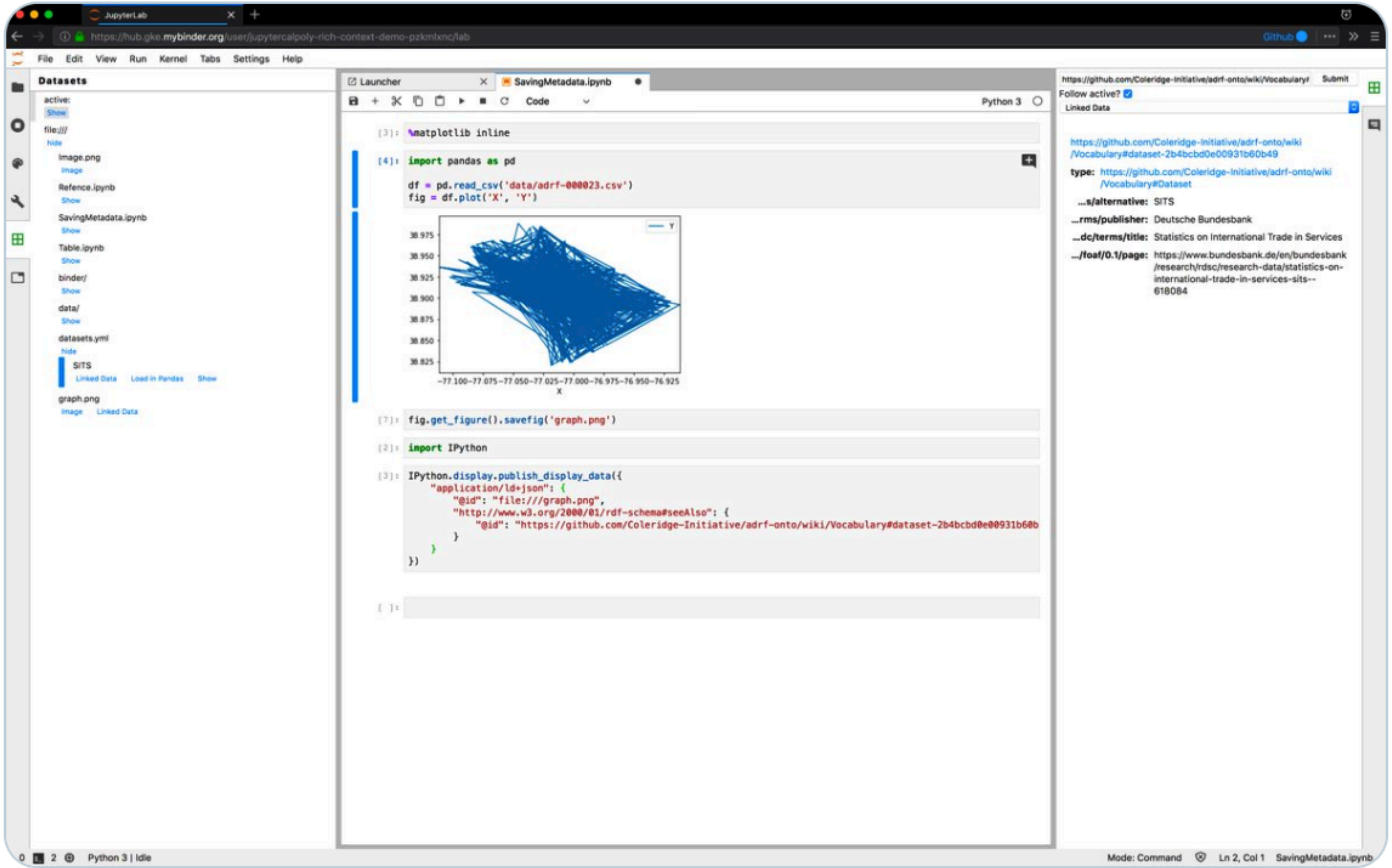
- JupyterLab **Commenting** and real-time collab similar to Google Docs
- JupyterLab **Data Explorer**: register datasets within research projects
- JupyterLab **Metadata Explorer**: browse metadata descriptions, get recommendations through knowledge graph inference (via extension)
- **Data Registry**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



```
[3]: %matplotlib inline

[4]: import pandas as pd

     df = pd.read_csv('data/adrf-000023.csv')
     fig = df.plot('X', 'Y')

[7]: fig.get_figure().savefig('graph.png')

[2]: import IPython

[3]: IPython.display.publish_display_data({
         "application/ld+json": {
             "@id": "file:///graph.png",
             "http://www.w3.org/2000/01/rdf-schema#seeAlso": {
                 "@id": "https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#dataset-2b4bcbd0e00931b60b
             }
         }
     })

[ ]:
```

https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#dataset-2b4bcbd0e00931b60b49

type: https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#Dataset

…s/alternative: SITS

…rms/publisher: Deutsche Bundesbank

…dc/terms/title: Statistics on International Trade in Services

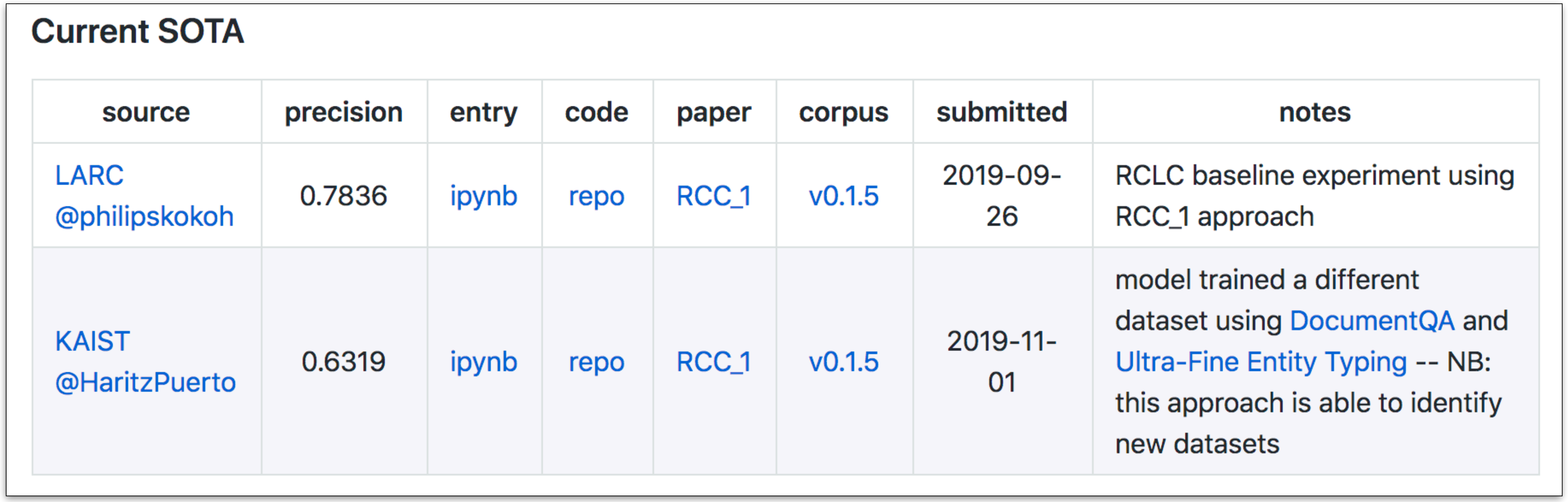
…/foaf/0.1/page: https://www.bundesbank.de/en/bundesbank/research/rdsc/research-data/statistics-on-international-trade-in-services-sits--618084

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

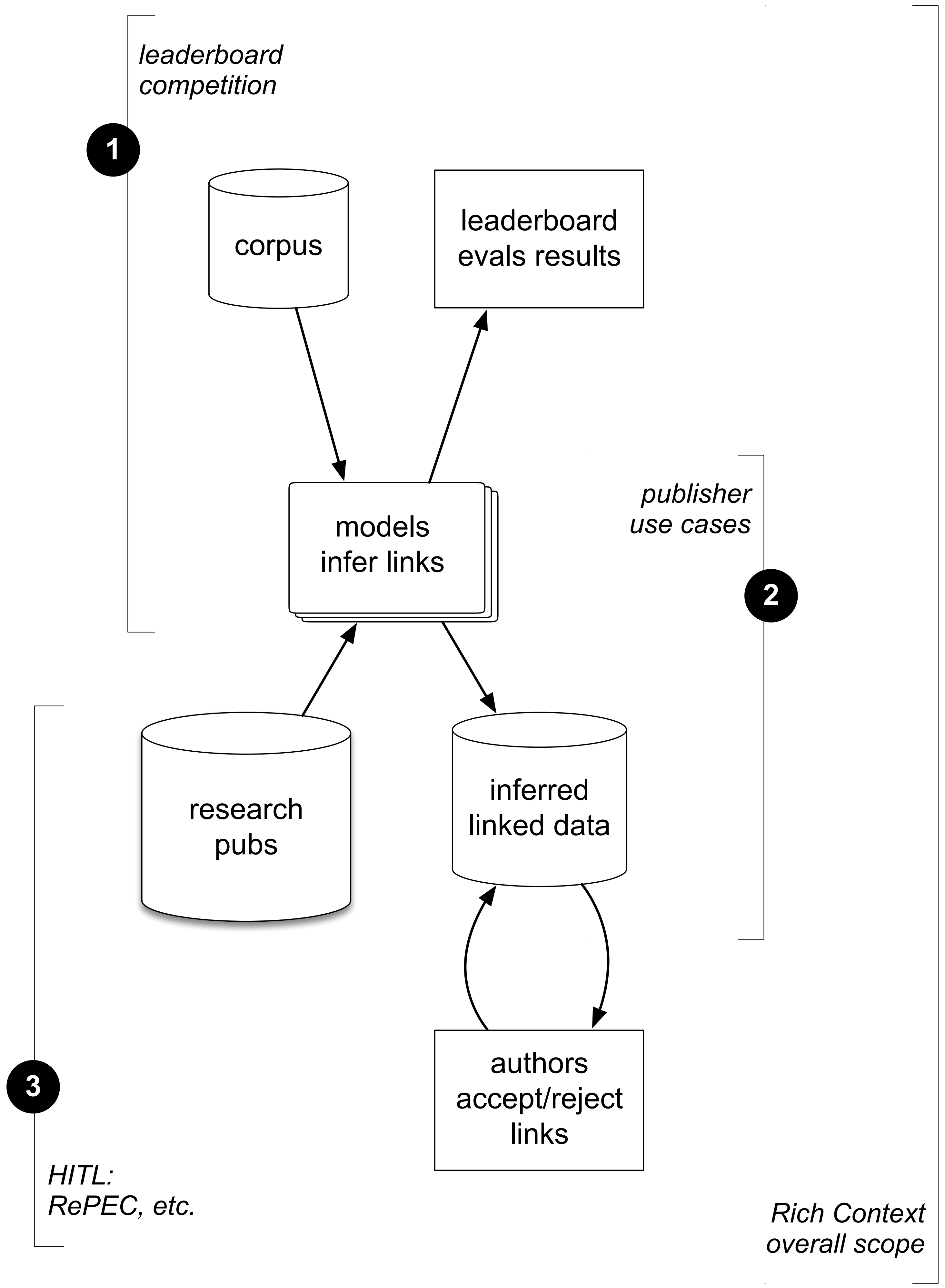


**Current SOTA**

| 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

- semi-supervised 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

# Additional Information

Rich Context @ NYU Coleridge Initiative
coleridgeinitiative.org/richcontext

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

EDITED BY
JULIA I. LANE • IAN MULVANY • PACO NATHAN

RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS

BUILDING THE NEXT GENERATION OF SCHOLARLY INFRASTRUCTURE

"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)