COLERIDGE
INITIATIVE

Rich Context Workshop

Summary Report

National Press Club
Washington, DC
Nov 15-16, 2019

Contents

Workshop Overview

The focus of the Rich Context workshop was to build a scientific basis for the empirical foundations of data science in government. Empirical research relies critically on knowing how data were produced and used before. Social scientists might define rich context as a dataset search and discovery process: what does the data **measure**, what **research** has been done by which **researchers**, with what **code**, and with what **results**. Computer scientists might define rich context as knowledge graph representation and recommender systems. Others might define rich context as promoting datasets to be a first class entity.

There now exists the technical capacity to build such a platform, as demonstrated by a successful recent competition https://coleridgeinitiative.org/richcontextcompetition. Computer scientists and domain experts in the life sciences have developed the scientific underpinnings necessary to build each component: document corpus development, ontology development for dataset entity classification, natural language processing and machine learning models for dataset entity extraction, graph models for improving search and discovery, telemetry to capture dataset engagement and use.

This technical capacity, combined with interest of funders in supporting data science, and the recent passage of the **Evidence-Based Policymaking Act** and the launch of the **federal data strategy**, made this an opportune time for such a workshop.

The outcomes of the workshop were intended to be:

1. A roadmap that will identify the current state, opportunities, gaps, and necessary investments.

2. The development of an interdisciplinary community of computer scientists, life scientists, and social scientists who can work together to address the problems.

3. The engagement of key stakeholders, notably funding agencies, and government agencies.

Workshop participants provided many suggestions and potential directions for moving Rich Context forward, which are summarized in this report. Notably, they identified three immediate activities, namely: (1) Developing a community of practice (2) Developing a “rich context” platform and (3) Specifying an initial use case with NOAA that can be tested for scalability across multiple agencies, starting with the USDA, a federal agency with deep experience with confidential microdata, and the Deutsche Bundesbank, which is leading an international effort to share knowledge about granular data (INEXDA).

Rich Context Workshop Summaries

Summary Sessions

Developing a Platform

Goal: develop a product that enables users to discover trustworthy datasets for research and commercial use cases.

This session built upon a previous session about the Amundsen platform from Lyft, for dataset discovery. That platform is based on knowledge graph of metadata about dataset usage. The attendees agreed that the platform should address two problems:

- Enable dataset discovery (and their related people, use cases, tools)
- Promote a trustworthy community of practice around that discovery through user experience

Consequently, the platform should focus on meeting customer-facing features and needs. The platform should inform customers what datasets are available, but should not provide data; rather, it should help customers engage with the data providers. Attendees agreed that the platform should be designed to enable customers to contribute to the platform and use it as a part of their daily work.

Action: The Minimum Viable Product feature set should be designed to ensure that no nodes or edges can be removed without breaking the ability to achieve the product goal (discovering trustworthy datasets). The desiderata are provided in the following table: a graphical version can be seen in the notes.

Minimal	**Aspirational**
• Dataset catalog • Datasets • Search – minimal • Data provider • Identity (hash of dataset and pointer to hash of data owner) • Community ◦ Trust metrics (minimal)	• Data catalog • Comprehensive – e.g. all agency datasets • Deep search - related datasets • Structured description - ontologies • Unstructured description • Time stamps • Contributor • Usage instructions • User feedback

Additional References:

Examples of Best Practices: IBM “AI Trust” initiative; Stratio AIDM (Rocket product)

Concerns: “Ground: A Data Context Service”, Joe Hellerstein, et al., CIDR 2017 http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf; “Datasheets for Datasets”, Timnit Gebru, et al., Microsoft Research 2019 https://arxiv.org/pdf/1803.09010.pdf

Building a Community

Goal: To develop a community that is self-sustaining and contributes to creating products that provide utility for a wide audience

Attendees noted that key features include not merely code commits but more importantly promotes the troubleshooting and understanding of use cases through community tutorials, mentoring, advocacy, meetups, and conferences. It is also important to set up a project that doesn’t cause its main contributors to burn out, since so much of community building involves coordinating with other organizations, having relationships with people who can resolve important blockers, and allowing the community to recognize itself. Participants also noted that humans make good “tidiers” in that they will clean up inaccuracies in existing content, but are typically not as willing to generate the original content.

Action: Develop project governance rules that work for multiple parties, e.g., GitHub contributor guidelines, which alleviate the problem that nearly killed the Hadoop open source project. Develop ways to arbitrate or some other tie-breaker clarified, to avoid factions and polarities that create governance deadlock.

Promising activities include

Build thematic organizations around data type:

+ Data catalogs. A version of {DataCite x Google Dataset Search}: a transparent public catalog of data, w/ source, date, and versioning; enough metadata to find what datasets exist related to a schema or concept; and a support-network honoring and recognizing people for making their data legible in this way.

+ Government data. There's space to help NOAA and other large dataset sources who have direct access to sensor-nets: feels like an explicit Underlay-registry use case.

+ Citation networks: Semantic Scholar, ResearchGate, PapersWithCode, and Primer.ai all want to contribute to a public citation graph. Name the package we're building together, iterate.

+ Wikidata feedback loops: Primer.ai for instance can fund young data scientists / residencies to organize [extracted] data that update Wikidata + WP. Needs a registry of "feedback claims" to hold onto them while they await community curation... and a community of data-fellows. Link w/ Schmidt fellows for data collaboratives?

* Possible partnerships

University collaboratives: A group of universities interested in solving the participation question: how should we all be contributing to a platform that comes out of this? Fitting this into university and department and library mandates.

Data broker collaboratives: addressing the public-private access question head-on: what this means for university research that wants to expand public knowledge while adhering to the norms and requirements of many non-public sources.

NOAA Use Case

Goal: Build an automated data inventory for NOAA that could be used as an ongoing use case for other agencies and users. This directly responds to the Federal Data Strategy: Action 7: Pilot an Automated Inventory Tool for Data.gov; it will also support efforts on Action 9: Improve Data Resources for AI Research and Development and Action 15: Identify Data Needs to Answer Key Agency Questions

Attendees noted that the use case should be of current interest, including information from multiple other agencies, not just NOAA. One possible focus would be on seawater inundation of coastal communities and tracing such effects as economic and social impact as well as emergency preparedness.

NOAA is a good use case because it can provide quick results. It already has multiple public inventories of their ~98K datasets, as well as highly sophisticated metadata management practices. What NOAA lacks is visibility into where and how their datasets get used, especially in commercial applications. When NOAA has that info available, their teams made very good use of it to establish feedback and improvements for customers.

Action:

Content: There is a set of ~20K research papers which are tied to NOAA datasets due to funding through NOAA; the NOAA library system will assist us to obtain those and related starting points. We know that ~80% of NOAA dataset usage is in China, i.e., those research papers are published in Mandarin; fortunately, the two top teams in the ML competition are based in East Asia. We can also leverage content from the ~1300 websites that NOAA maintains to describe their datasets, i.e., for training the ML models to identify datasets in papers.

Technical: A service such as DiffBot could be used to find NOAA dataset references in usage, or to expand the "lookalike" analysis (apply ML models) to find more use cases after starting from an initial seed set.

User Interface: The focus in the automated data inventory for NOAA would be on:

- people: build map of people and their expertise who leverage NOAA dataset – which can be extended via author co-occurrence analysis
- use cases: how the datasets get used, which is often surprising and may shape the course of future work by NOAA (e.g., migratory bird studies as a surprise use case).

- tools: what kinds of tools (e.g., ML algorithms, other data integrations) commonly get used with the NOAA datasets

References: NCA4 https://www.globalchange.gov/nca4 sets a benchmark for the NOAA goal of success in this project; the **Arctic Report Card** https://arctic.noaa.gov/Report-Card is another good example (2019 report is coming out soon).

Structured Sessions

Use Case: Deutsche Bundesbank

Goal: Identify what are we doing now that would be transformed by access to dataset search and discovery tools, and what new research could we do if we had the right tools and data? How can tools be calibrated - by application need, user ability and resource constraints?

Society and the economy is extremely complex; data need to be integrated from multiple sources to promote better understanding. Attendees noted that access to data would permit studies to be replicated, especially with proprietary or restricted data. Finding new data requires more standards and schemas. They noted it would be best to start with 15-20 datasets, rather than boil the ocean.

Action: Build a Worldwide inventory of datasets including relations and linkages to other datasets. Develop a *real* data standard. Standards need to be extremely explicit – medical data exchange example where several people are compliant, but still can't talk to each other.

NLP Research

Goal: Tackle problems in entity linking and coreference. Identify is the state-of-the-art (SOTA), given so much progress with transformers.

Attendees noted that NLP SOTA has advanced with the use of BERT and XLNet methodologies. The models rely on transformers which use a mechanism called a tension to map correlations across an entire sequence simultaneously. A major challenge isn't misclassification, but missing things - silent failure - where a model just does not find something that it should. Understanding causation is critical, because so many efforts are black box solutions; clarity of cause is often missing This lack can introduce bias, and it is important to have diverse, challenging, adversarial examples across any area of interest together with governance practices for versioning of models and data that trained them.

Useful reference: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/

http://nlpprogress.com/english/entity_linking.html

Action: Begin to scale out workers for specific Rich Context tasks. Understand biases. Develop discussion about governance about the model itself, and what privacy issues might emerge.

Metadata Exchange

Goal: Harmonize metadata across various sources, so that data can be consumable by stakeholders. Metadata standards must guarantee consistency and durability. A lack of persistent identifiers and schema make metadata exchanges challenging,

Attendees noted that there are lessons to be learned from failure: the Cancer bioinformatic grid (caBIG), which was a project for secure transfer of cancer data for research. It required data providers to submit metadata which limited contribution. Didn't make use of semantics, there was a heavy burden on upfront annotation. The ambition was to have deeply annotated data, but they didn't budget for it and was a disaster.

Attendees noted that data sharing and data purposing is more a social than technical problem, so the strategy might well be to focus on identifiers, and finding incentives for having people use the existing frameworks. While persistent identifiers are important, standards must be flexible enough to permit change over time.

Since much of the motivation here is ultimately about ways to reproduce research (or recommend resources for similar research) one of the better examples to consider closely is PapersWithCode https://paperswithcode.com/sota in which each unit provides: identified use cases. consensus performance metrics, a public training dataset use for benchmarks, performance baselines and links to code and papers for top performers. This approach allows ML researchers to compare other results, and what how other teams perform over time – a dynamic picture of best performance in the world for a specific research area. Note that data sharing in PapersWithCode is more of a social mechanism than a technical one.

Action: Work toward an authoritative source of reconciling and disambiguation. Consider the potential for Wikidata being the place to collect metadata. Build on paperswithcode/sota as a possible approach. Build a reputation mechanic that ensures that asking AND answering questions well is part of how reputation should be scored. The mechanic could also incentivize reviewing

Useful Reference: https://paperswithcode.com/sota

Human-in-the-loop

Goal: What are the touch-points for humans involved in machine learning end-to-end workflows. How can we build those touch-points into the ways that we use datasets through policies, data collection and reuse of datasets?.

Attendees noted that Human-in-the-loop (HITL) is a broadly used term with several different meanings or connotations. In general the term means adding people into feedback loops involving decisions made with data or outputs from machine learning models. This practice is also widely used to prioritize where to add labels (typically an expensive task) to unlabeled datasets, so that they can be used to train ML models. When combined with ML model explainability, HITL can provide a "two-way street" where human experts help train ML models, while ML models help aggregate the organizations learnings from the people involved. Attendees noted the importance of considering the full ML lifecycle of Data Prep -> Feature Selection -> Train Models -> Evaluate Models -> Use Cases, and all of the ways that people have crucial inputs at those points.

Note that virtually every dataset has some component of human input: decisions about data collection, rules about data governance, etc., which must be guided by policy and legal frameworks.

Action: Define which humans and which feedback loops. Bring organizations to think about their workflows rather than fully automate them. Increase focus on the failures resulting from skills getting automated – without educated intervention, models are fractured and "poisoned".

Useful references: “Active Learning Use Cases" https://derwen.ai/s/d8b7; Snorkel https://www.snorkel.org/ (weak supervision)

Knowledge Graph

Goal: Define a knowledge graph and the value for rich context.

Attendees noted that a knowledge graph refers to a graph of entities and relationships within a domain - where data and semantics are combined in a graph structure to represent knowledge. Examples include ORCID - person identifiers are connected to research works. Their graph is made possible and sustainable by a unique. Wikipedia is an example of a user-built knowledge graph. RDF (Research Description Framework) ontologies can be used to define and model relations and entities although there are limitations. When building a KG, design is informed by defining semantics and metadata in the KG. RDF/graph stores can be stored in databases like Amazon Neptune, GraphDB, and queried by SPARQL.

Is RDF right for Rich Context? The benefit is that RDF handles identity, provenance and semantics. But RDF takes effort to maintain, and few large industries use it - they have their own home-grown ontologies and vocabularies. A best practice is to reuse specific vocabulary and have consistency within the graph, as well as to develop or acquire tools that integrate multiple representation schemes – particularly the ones we care about.

Many experienced attendees warned strongly against using RDF.

Action Need to distinguish which languages/ontologies to use, since it is difficult to motivate people to switch away from their favorite. Define methods for documented, maintained metadata. Develop consistent ontologies. Need guidelines for reuse. Consider Wikidata as general go-to publishing mode

Resources: https://cacm.acm.org/magazines/2019/8/238342-industry-scale-knowledge-graphs/fulltext

Centralized Services

Goal: Identify where we need centralized services, nationally or globally, to augment what’s missing now, e.g., a global repository of datasets, with persistent identifiers. Define a “service overlay” on the knowledge graph to complete the knowledge lifecycle

The group worked to identify where a centralized service for Rich Context would be useful, and define a ‘service overlay’ on the Rich Context Knowledge Graph. A centralized service would augment what’s missing now, e.g. data discovery tools, measures of data quality and a global repository of datasets with persistent identifiers. A centralized service would need to provide data in a way that is easy to integrate into multiple existing infrastructures/spaces - some which are well defined and mature (e.g. scholarly infrastructure), others less so (e.g. NLP). Publishers have an opportunity to have better metadata and have content that is really well tagged, making a centralized service where that data can come from a really appealing prospect.

The key question is how to manage trusted access and make use of standards like DDI. ISO 19115 and Dublin Core metadata. The challenges associated with delivering Rich Context metadata through a centralized service include: privacy, security and access concerns when joining datasets. Uptake in participation and contribution back to such a service is likely driven by an egocentric mechanism. With the ReplicationWiki, uptake in participation increased when research was linked to RePEc.

Action: Work towards defining a "related data" metadata schema, Check the possibility of using existing standards, which may or may not have coverage for the use cases we've heard Investigate models of automating access to data that combines permissions schema (inspiration might be IAM system from AWS).

Useful examples: Papers with code ; Underlay (KFG); World Bank (World development indicators); Replication Wiki ; Semantic Scholar ; Google data search Bundesbank

Using INEXDA standard; US Statistical agencies

Scholarly Infrastructure

Goal: Identify how metadata resulting from Rich Context can augment the existing scholarly infrastructure

Attendees noted that Integrating the outcomes from Rich Context into existing scholarly infrastructure would benefit multiple stakeholders. It would reduce the cost for researchers in searching through the complexity inherent in the explosion of research objects. It would help data providers show the value of datasets and enable monitoring of secondary usage of data. And it would help the scientific community more broadly to facilitate transparency and reproducibility of research.

The challenges to be faced include: inconsistent vocabulary, lack of persistent identifiers, license restrictions, limited personnel and staffing, the difficulty of providing search across multiple sources of infrastructure and lack of consistent method of dataset reference

Standards for metadata usage emerge over time, but **they take time to form a community of practice**. There are trade-offs regarding whether to establish standards, especially when attempting to establish standards too early.

Action: Develop good use cases with the following features: good data models for integration with a clear scope, engaged infrastructure providers willing to add in rich context hooks, an engaged community willing to test out ideas, an immediate value proposition resulting from connecting people and projects, a demonstrated impact on the community and a system that's extensible to accommodate new use cases

For a listing of scholarly infrastructure and discovery services, see "Scholarly Infrastructure" on the workshop wiki.

Privacy Vs. Utility

Goal: Access Reinvented. Understand risk vs benefit to optimize access to restricted data

The challenge is to quantify the impact and measure the social benefit of data collection to expand access to restricted information.

Data collected by the Government under a pledge of confidentiality could have significantly more value and social benefit if it was easier to access and use. Our digital society is rapidly changing and timely access to data that can be used to make a difference in people's lives is critically needed. Timeliness is critical, especially among vulnerable populations, where patterns shift and data has been difficult to collect. The geographic location and response has to be completed in a meaningful timeframe to have any effect, or the chance might be lost. The problem is when access, research, and evidence building delays take too long to produce meaningful output.

The current system has been locked into a bimodal 'public or restricted' only paradigm. Data in the 'restricted' space is plagued by significant administrative costs and delays to get access. Technologies like Differential Privacy are not mature enough to support the social sciences where longitudinal studies and linkages across board datasets of varying quality are needed. Synthetic data sets with validation/verification servers are expensive to create and take too long to produce such that broader use as a solution is untenable.

Consider the wide variety of personal information already collected and available by commercial companies, not to mention the billions of records exposed through data breaches in 2019 alone. Given these conditions, how do we ensure trust in government with respect to collection of personal microdata, and determining the tradeoff between confidentiality and utility? Such a large focus has been on protecting the data that many opportunities are missed to optimize the value of data collection and use.

Creating other tiers or modes of access based on balancing risk versus benefit is difficult because the consequences or larger benefit of research are not well measured and defined. We do know that access makes a difference and one example is Heidi Williams' paper of Celera, in which she compared genetic update from Human genome project and private collection. Compared to open access, Celera was suppressing uptake/benefits by ~30% by limiting access to the data.

Actions: Use rich context to measure the social benefit of data collection. Work within the statistical community to use these measures when evaluating the risk versus benefit tradeoffs to expand access to other access tiers outside the traditional bimodal system of public-restricted. Leverage more public-private partnerships to gain timely access to data that can be used more broadly without as many restrictions. To assist in gathering evidence, conduct a pilot to emulate the European method of opinion research measuring the trust/benefit of data collected from various sources.

Resources: Intellectual Property Rights and Innovation: Evidence from the Human Genome (Heidi Williams): https://www.nber.org/papers/w16213

Lane, Julia, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, eds. Privacy, big data, and the public good: Frameworks for engagement. Cambridge University Press, 2014.

Catalyzing a Community

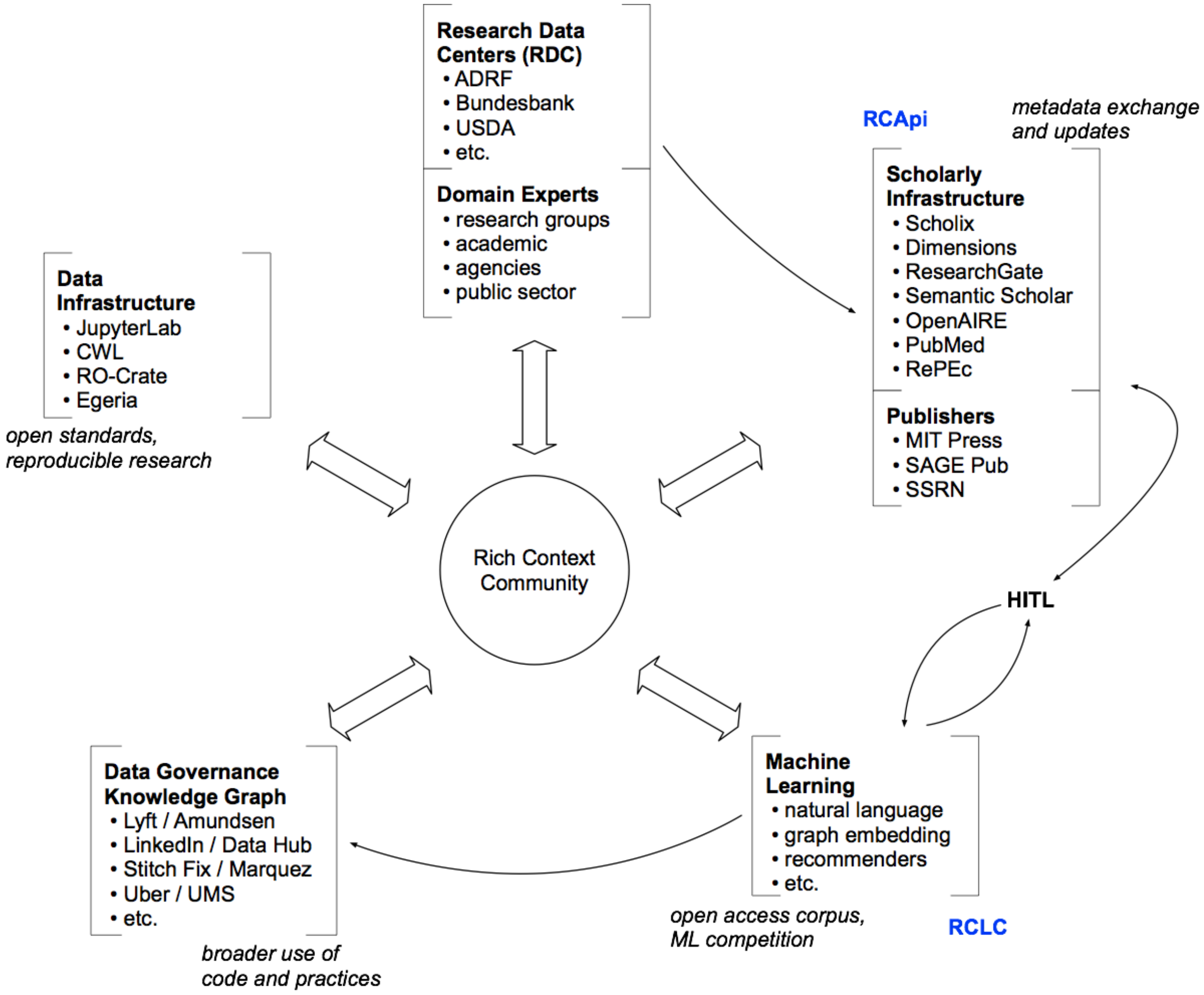
Goal: move toward coalescing sources and promoting widespread collaboration. To do so, user communities must be defined and identified around a common need that is well-scoped.

Attendees noted that a community can be identified through common links that are generated by common problems; it should be small enough to feel a sense of belonging, but big enough for wider-scale contribution opportunities, with an overlap of users and contributors. Wikipedia and Stack Overflow have been successful because whatever the “thing” that brings a community together solves a problem well - the common motivation behind these actively contributing communities is that there is a common, widespread need, and easily accessed usability value. Stack has many separate sub communities (engineer, stats, data sci), all with distinct pages/titles and centered around the use of the same programs. It’s a tool that is in effect, a facilitator. Wikipedia fit a need that people had, but Britannica refused to address it. Wikipedia is more recent, not as precise, and the result was years of stigma. However, by contrast it offered a quick turn-around of articles and “out-Agiled” Britannica because it understood that the user requirement was not to have a perfect product, just to have “something”. Users were motivated to contribute because it spoke to a common/widespread need and was freeform.

Community building is hard in the rich context use case because we have a range of participants with different needs, all who have a stake in what we are talking about. In terms of incentives, the thing that brings a community together is also what tends to define it In the realm of metadata, the goal in developing and catalyzing a community is to narrow the scope in order to meet the needs of communities (or groups thereof), rather than the needs of “everybody”. The place to focus initial efforts is on the smaller-scale, in identifying datasets with a large audience, user base, with widespread application. An example of this is NOAA weather data, which is used throughout government as well as the private sector (e.g. Lyft). These data are ingested by many third parties, and eventually relayed to users - it is applicable to so many domains, and everyone knows what it is

Action: Provide for common needs in data infrastructure and metadata exchange. Create flexible frameworks and methods to evaluate consistency/performance. Guarantee that as needs arise to modify, modifications are uniformly made.   Integrate across existing popular open source projects, in lieu of creating standalone projects. Consider Dependencies: JupyterLab, spaCy, PyTorch rdflib, etc. Standards: Egeria, W3C standards for metadata. Make use of existing frameworks: Amundsen and its emerging category. Develop approaches for evangelizing these projects together. Provide content through AI metadata generators and enable community to “tidy up”. Incentivize with rewards

Useful reference list of involved communities



Business Models

Goal: Define business models that fit the problem and the community, and identify how to seed initial work with research-funding support, and show how to become self-sustaining

Attendees noted that the challenge with not for profit activities is that philanthropic funding takes the place of venture capital funding. But while venture capital provides both funding and business help, philanthropic funding does not provide the latter. If rich context is considered to be a product, it can be seen to be more sustainable, since business cases can be constructed around products and services for the people who might pay you. The focus should be: which activities can have immediate value to the financial organizations which then lead to sustainable income over the next four or five years? Data have been challenging to fund, since they have some of the features of public goods (excludability and rivalrousness). Because data access, particularly for confidential data, have some rivalrous aspects, they are common pool resources: governance is therefore critical.

Attendees wondered whether a centralized entity was necessary, or whether a set of standards and their related ecosystem would suffice.

Attendees identified the following possible sets of products and Services that rich context could provide: persistent identifiers; service to inform universities & researchers when others are using their data/research; core software libraries; service to identity provenance; data inventories; stewardship-as-a-Service; data quality evaluator.

Action: Build a business plan from the beginning. Focus carefully on the costs.   Be careful about too much dependency on a small group of people who are particularly passionate and invested. Get business consulting help

Useful examples: ORCID, DRYAD, REPEC, CrossREF, DeepSpace@MIT, PLOS, ARXIV

Unstructured Sessions

Dataset Search

Goal: when you need to search for data, what's your query? What are the UX personae for dataset search?

Attendees discussed the definition of a dataset. CERN has formal definitions of what is a dataset and differentiate between types of datasets in different contexts. Do we need a standard definition of dataset when we can provide useful tools without strict consensus on the meaning? RePEC has been working for several years to find common metadata basis for describing datasets. Definition for 4 typographies: (1) raw data (2) events of interest (3) more processing (4) even more processing. Issue is these tend to be siloed by discipline.

It is critical to understand how people search for data: by keyword, for similar data, for similar purposes (effects/results/performance metrics). Also why people search for data: do they look for data by first asking questions; do they want to try to use the same data and compare to existing results, do they want to put their own data together and need to know what data are needed so look at studies that have been done and see what is necessary to do something similar. In all cases, there’s a connection to the relevant literature.

Attendees noted that the problem of attribution/citation should be isolated from search and discovery. How can we surface signals of researchers finding that datasets are bad, not useful -- warn others -- could Jupyter telemetry features help here? We need a schema for feedback (airport feedback as an example) which should go back to the data provider - pre-processing and addressing quality issues should be the responsibility of the data provider. In the other direction, a mechanism for providers to warn users about known issues. As datasets get distributed by Google, Amazon, etc, then the direct feedback becomes even more removed. Are there means for data consumers to help each other (similar to StackOverflow)? A strong feedback loop is needed between data user and provider.

Papers that reference data with proper attribution tend to be more highly cited. There's "data rot" where universities control the data access (with Data Use Agreements) but not the maintenance of making the datasets available in an ongoing basis -- longevity is a problem. Government-sponsored research requires that datasets go to repositories that have maintenance policies. NIH rules state that your data comes off embargo after your first publication, so researchers can’t hoard datasets to avoid competition. (Deborah McGuinness, RPI). Research also must state any foreign dependencies. Impact metrics on a dataset (e.g., second-order citation graph) of what a dataset enables, that provides decision makers with "who to reward".

The Making Data Count project has been researching data usage, data citation, and the reach of data for the last six years and while infrastructure is ready for capturing both usage and citation, there is some more community work to do.

Key Takeaways

1. having no metrics is bad, having one metric is bad, a combination of poor-ish metrics is not better, having bad metrics is dangerous

2. behavior of dataset search has changed: more search engine referrals, hence the dataset providers much change how they track and understand the usage
3. needs to be more feedback loop between the data consumers and the data producers

Useful references

- https://blog.datacite.org/why-data-citation-matters/
- UMETRICS project (iris.isr.umich.edu)
- "Responsible Use of Metrics" https://zenodo.org/record/3507812#.Xc7uSFdKhPY
- Open Data Metrics “Lighting the Fire” Lowenberg, Daniella; Chodacki, John; Fenner, Martin; Kemp, Jennifer; Jones, Matthew B. http://opendatametrics.org

Cloud Computing

AWS is building a pipeline for end-to-end computing from data ingestion to processing and machine learning. The pipeline is a combination of infrastructure, tools and processes that pulls best practices from various ecosystems. These are pseudo automated tools that organizations can customize with the tooling of their choice. The goal is to shorten the process through which people develop their capabilities. Data goes on a journey from ingestions, to storage, to analysis and finally insights.

The pipeline focuses on having open-source libraries and tooling. The pipeline has an opinionated set of open source tools that would be common to many domains. E.g if you take TensorFlow, and build a software development kit on the cloud to launch a TensorFlow job. An ‘opinion’ is the approach made available by the kit.

Other pipeline features: Tools to improve/filter a dataset to a usable meaningful subset of it; The end-to-end solution enables you to track the output upstream to help you tune a model, figure out when you retrain it; AWS data exchange - data store with metadata and field descriptions etc

Attendees questioned the bounds of end-to-end and the structure of moving documentation along that workflow given exogenous information and the challenges of getting workflow metadata to be interoperable across other vendor platforms

Key Takeaways

1. Metadata on workflow would come out of an end-to-end cloud computing pipeline
2. Linkages between code bases and datasets are important for tracking versioning. That information needs to be readily available to an analyst to prevent error/inaccuracy or misunderstanding of what is in the dataset

Useful references

The tools exist to create knowledge graph from the relationship between data and the code bases that transform them. https://aws.amazon.com/sagemaker/

Wikileaks

Goal: There is no canonical version of leaked data, and we don’t have community norms to drive leaking. In the scientific community, there is a cost for leaking data (in gov’t, penalties). How do we respond morally to leaked personal data?

Attendees noted that it is very hard to have public trust in analyses of huge amounts of unstructured, person-level data. Is it possible to quantify the impact from leaks? Wikileaks speaks to the question of whether it is possible to solely rely on trusted sources to hold and analyze data. How could deep-fake versions of data be identified, since adversarial data injection (data deep fake) seems inevitable. Everything that is done to make it easier to discover and search data also makes it easier to leak.

Will we go back to old reputation signals--where people just trust someone who's built a trusted reputation? Will data become just like other information sources--possible to manipulate, and therefore harder to verify or trust?

Key Takeaways

1. Need for an organization that connects data scientists and journalists, so that journalists can get access to good analysis of large data sets. Hashing to confirm data provenance.

2. Develop community norms, protocols, codes of conduct for journalists and data scientists who are working on leaks together

Data Provenance

Goal: When seeking a dataset, users should know who created and curated the data as well as the context of the data.

Attendees noted that there is great interest in better understanding data rights and ownership and how people, including indigenous peoples, can use datasets (the FAIR principles - there is also CARE - Collective benefit, Authority to control, Responsibility, and Ethics). The general rubric of governance, and indexing specific datasets at a time doesn’t give a picture of usage. Notice and informed consent is insufficient to deal with the later uses. Institutional mechanisms must be in place for both individual data and tracking data once it has been aggregated and governing usage. NIH is trying to enable pooling of the data so more meaningful conclusions can be made from the data. But if you don’t know if columns are the same, the meaning of the columns are obscured. There has been some work done in copyright - TK labels (https://localcontexts.org/tk-labels/).

There has been a lot of work in digital rights management which is controversial but a useful part of the conversation. Semantic scholar has been working on citation ‘intent’ classification - intent as to why another paper was cited. Was it cited as background, methods, results – and is it possible to extract information about the kind of claim someone is making in a paper.

Attendees questioned the incentives to go back and update datasets for usages that were not thought of at first. IRBs and data management plans are insufficient. Researchers should be required to speak with their institutional libraries to make data management plans as proposals are awarded - the library systems are disconnected from researchers in the planning stages (.

Researchers ask about the contents of the datasets before actually requesting the dataset - the "beacon" mechanism could be really powerful. Nature requested data availability statements. 92% said data available upon request (eg. not available).

Key Takeaways

1. Bring up ethics of data usage around the table
2. Governance of usage is as important as datasets, the models and derivative datasets and other products from the datasets.
3. Data provenance is not just a static documentation of who, what, when, where, and how. Rich context should also capture dynamic human-centered activities.
4. Data studies vs data science

Resources

https://www.biorxiv.org/content/biorxiv/early/2018/10/17/443499.full.pdf

https://www.sciencedirect.com/science/article/pii/S1532046417300990

https://journals.sagepub.com/doi/abs/10.1177/0340035216682041?journalCode=iflb

https://localcontexts.org/tk-labels/

https://en.wikipedia.org/wiki/FAIR_data

https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/

https://beacon-network.org/#/

https://brave.com/

https://beacon-network.org/#/

Space studies board https://sites.nationalacademies.org/SSB/index.htm

Human centered data science: https://depts.washington.edu/hdsl/

https://brave.com/

Problems with classifying citations, and data citations; includes refs to early attempts to classify. Chapter in Chris Borgman's 2015 book, and developed also in this edited book, providing original and open access references: Borgman, C. L. (2016a). Data citation as a bibliometric oxymoron. Retrieved from https://escholarship.org/uc/item/8w36p9zf

Human in the Loop

Goal: Iteratively involve the human aspect and response to data development. Humans are the stakeholders from data collection to consumption.

Attendees noted that one concern with keeping humans in the loop with respect to data approaches involves cybersecurity, and the development of systems that take into account all the people involved in the process, which works from all sides (Which humans? Which loops?). So in some cases, human out of the loop might be more appropriate.

There is a movement to automate more processes, relating to the idea of provenance and how we can lose sight of the chain of control and ownership unless it is well documented and understood. In moving from the old model of 100% manual data collection to the current methods of people feeding in data, often unknowingly (via apps, maps, etc.), there are pros and cons. An important pro is that automation greatly reduces technical errors, while a con is that the algorithms automation is based upon are subjective in respect to human biases.

Attendees discussed the question of which point it is appropriate to insert human intervention into the process. What are the places in research where humans especially need to be involved so that results that are not very biased or outright wrong. However, as we are on the forefront of this issue, it is hard to identify the specific point at which intervention needs to occur.

The conversation is centered around metadata features (and needing standard ontology for its effectiveness), and on potential policy outcomes (implicit knowledge and opinions of researchers is captured in analyses). The need to identify how to find datasets, and through connecting dots, how to best link it to relevant research and its context in turn creates another feedback loop in the human in the loop element of the Rich Context process.

Resources:

Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation: https://zenodo.org/record/1406002#.XdLEFTJKjOS (although the implementation has been challenging)

Jupyter

Goal: Update on relationship with Rich Context

JupyterLab Metadata is the exploring arm of Rich Context. It links with existing data catalogs and populates semantic links between different types of knowledge (ie: that stored in datasets, notebooks, individuals minds, publications, books, etc.)

JupyterLab Commenting is the communicating arm of Rich Context. Using JupyterLab commenting a single user can currently take notes and track progress related to work in notebooks, or any activity you can launch in JupyterLab.

The Jupyter Telemetry System enables administrators to automatically collect structured events, such as user and server actions, from Jupyter applications. It serves a number of purposes: i) operational and security monitoring, ii) intrusion detection, iii) compliance auditing, and iv) a-posteriori analysis of the platform's usage and design.

The team is working on a secure multi-server environment (that has a single canonical version of the notebook) that will likely mostly be used with Jupyter Hub. It will allow admins to build knowledge graph from user interaction

User feature/UI requests include: simple string search; copy code snippets; bake filetype metadata (png, pdf, etc) into metadata; "publish my data/metadata"; include requirements file

(info on libraries needed to run a notebook)

In process: ability to pass long/share notebooks within a Hub. "Notebooks are not the shareable unit; directory is"; commenting

The goal is to build a community of integrations; write plugins for particular datasets or datatypes. The team expects to find that there are clusters of plugins that are useful for particular domains or types of data. There is a nascent marketplace of plugins? Notebook will have JSON-LD built in.

Lightning talks

Powerpoints available here

Posters

Posters available here

Additional Resources

http://18.216.198.231/wiki/index.php?n=Main.Resources

Password: available upon request to dataanalytics@coleridgeinitiative.org

Detailed Workshop Goals

Goal #1 -- Identify compelling use cases that would be transformed by access to dataset search and discovery tools (starting from Evidence-Based Policymaking).

Goal #2 -- Take stock of existing practices and identify the gaps for the following:

* Goal #2.1: entity linking and co-reference (ML/NLP research);

* Goal #2.2: metadata exchange: persistent identifiers, crosswalks, data dictionary discovery, translation between W3C => ISO, etc.;

* Goal #2.3: knowledge graph representation and inference;

* Goal #2.4: how the resulting metadata can be used to augment scholarly infrastructure (SAGE Pub, MIT Press, RePEc, ResearchGate, etc.);

* Goal #2.5: human-in-the-loop approaches: semi-supervised learning, weak supervision, and other variations; plus, how to incorporate authorized contributors.

Goal #3 -- Catalyze a community that works together to integrate open source projects for common needs in data/metadata infrastructure (JupyterLab, spaCy, PyTorch rdflib, Egeria, W3C standards for metadata, Amundsen and its emerging category, etc.).

Goal #4 -- Identify where we need wider use of metadata exchange (e.g., persistent identifiers, updating publication metadata, etc.) to complete the knowledge lifecycle.

Goal #5 -- Define a platform (akin to Amazon, Etsy, LinkedIn) for the initial use cases, which can be broadly adopted:

* What are the desiderata?;

* What does an implementation look like?;

* How readily can that be implemented?

Goal #6 -- Generate business model(s) that can be seeded with initial research-funding support and subsequently become self-sustaining.

Workshop Agenda

Event Schedule

The Rich Context Workshop offers a mix of Lightning Talks, participant-determined Unconference Sessions, and Structured Sessions on relevant topics. Schedule-to-date is below (and subject to change).

Friday

08:30 -- Doors Open for Check-In

09:00 -- Opening Session

09:45 -- Lightning Talks (First Amendment Lounge)

1. The Federal Data Strategy, Evidence-Based Policy, and Rich Context
   Speakers: Nancy Potok, Chief Statistician of the United States & Margie Graves, Office of Federal CIO
   Time: 09:45
2. Lyft's Data Discovery and Metadata Engine - Amundsen
   Speaker: Mark Grover, Lyft
   Time: 09:55
   - The problem of data discovery.
   - Evaluation of solutions and the decision to build.
   - Open source Amundsen project and how it models data sets, queries, etc.
   - The Future: Building other applications on top of the same metadata.
3. Persistent Identifiers
   Speaker: Jo McEntyre, EMBL-EBI, EuropePMC, Project Freya
   Time: 10:00
   - This talk will review the landscape of available persistent identifiers for different parts of in the research infrastructure, discuss the maturity and uses of various types of identifier and demonstrate the power of identifiers when used in combination.
4. Enhancing data-informed emerging technology (AI) policy analysis
   Speaker: Dewey Murdick, Georgetown
   Time: 10:05
   - This talk will quickly summarize the Center for Security and Emerging Technology's (CSET) efforts to acquire and enhance data so that it can be used to inform emerging technology policy discussions.
   - The substantial investment in data acquisition (licensed and otherwise), fusion, quality control, and the manual annotation of data sets for the training of useful language models will be of particular focus.

   - The audience will be introduced to our efforts and will learn of opportunities to collaborate on selected projects.

5. Digital Government and EITC
   Speaker: Sue Marquez, Rockefeller Foundation
   Time: 10:10
   - The Rockefeller Foundation is currently working with organizations that help low-income folks access government services more effectively, especially focused on the tax incentives and processes around the Earned-Income Tax Credit.

6. Contextual Label Smoothing with a Phylogenetic Tree and an Airplane-mode Demo on an S10+
   Speaker: John Kaufhold, Deep Learning Analytics at General Dynamics Mission Systems
   Time: 10:15
   - Recognizing the species in a photo of a living thing is a long-tailed challenge that stresses learning species categories with very few labeled training examples. On the global iNaturalist fine-grained visual categorization challenges in 2018 and 2019, both aimed at benchmarking the state-of-the-art in this speciesID challenge, the Deep Learning Analytics Team placed first in the United States, outcompeting Baidu in 2018, and Facebook's FixNet in 2019.
   - Deep Learning Analytics' entries took a judicious approach to label smoothing (called "Contextual Label Smoothing" in our 2019 paper) guided by a SME-provided phylogenetic tree as context.
   - To sketch the main novel idea, we use proximity on the phylogenetic tree to not label smooth a training example of a humpback whale to learn a dissimilar (i.e. distant on the phylogenetic tree) monarch butterfly category, but we do label smooth a training example of a gluphisia moth to have a nonzero contribution to learning the similar (i.e. close on the phylogenetic tree) monarch butterfly category.
   - We will also demonstrate this network running in real time on a Samsung S10+'s DSP in airplane mode.

10:30 -- Break

11:00 -- Unconference Session

12:00 -- Lunch

12:30 -- Unconference Session

13:30 -- Lightning Talks (Room: First Amendment Lounge)

1. Data, Unstable in Concept and Context
   Speaker: Christine L. Borgman, UCLA Information Studies
   Time: 13:30

- Data exist in the eye of the beholder. One person's signal is another's noise. While metadata and ontologies are necessary for data to be FAIR (findable, accessible, interoperable, and reusable), they are rarely sufficient. The context of data evolves over time, which influences data practices and management in complex ways.

2. Enabling Scientific Discovery Through Machine Learning of Large Astronomical Data
   Speaker: Tuan Do, UCLA
   Time: 13:35

   - Very large multi-dimensional datasets are now available or will very soon be available in astrophysics. Scientific discovery will rely more and more on our ability to understand these data.
   - Discussion on how astronomers are currently using machine learning to help enable discoveries as well as current limitations on using these methods.

3. Medical Surveillance
   Speaker: Ophir Frieder, Georgetown
   Time: 13:40

   - This is a trailing part of an invited talk presented to a group of social scientists; the audience was excited by / horrified from / and annoyed with the realities presented.

4. Moving Off Record
   Speaker: Euan Adie, overton.io
   Time: 13:45

   - I have been studying grey literature sources (think tank research, government docs, reports from foundations etc.) and they are missing many of the features we take for granted when doing text mining using with the scholarly record.
   - Outline of the biggest issues to see if anybody else has solutions or the same problem with other types of content.

5. The Case for an Institutional Consortium
   Speakers: Amy Brand, MIT Press & Sam Klein, MIT Underlay Project
   Time: 13:50

   - The MIT Knowledge Futures Group aims to accelerate the path from research breakthrough to societal benefit; develop and scale technologies that open, enrich, and fortify our knowledge infrastructure; and galvanize a movement towards greater institutional and public investment and participation in that infrastructure.
   - This talk will focus on the KFG's efforts to create a multi-institution consortium.

14:00 -- Structured Sessions

1. Use Cases (starting from Evidence-Based Policymaking)
   Session led by: Stefan Bender, Deutsche Bundesbank
   Room: Murrow
   - What are we doing now that would be transformed by access to dataset search and discovery tools?
   - What new research could we do if we had the right tools and data?
2. NLP Research
   Session led by: John Bohannon, Primer AI
   Room: White
   - Tackling problems in entity linking and coreference.
   - What is the state-of-the-art (SOTA), given so much progress with transformers, etc.
3. Metadata Exchange
   Session led by: Jo McEntyre, EMBL-EBI, EuropePMC, Project Freya
   Room: Lisagor
   - Persistent identifiers, crosswalks, data dictionary discovery.
   - Translation between W3C, ISO, and other standards.
4. Human-in-the-Loop Approaches
   Session led by: Paco Nathan, Derwen AI, NYU
   Room: Bloomberg
   - Semi-supervised learning, weak supervision, other variations.
   - How to incorporate authorized contributors.
   - Design thinking on behalf of the people involved in these systems.

15:00 -- Structured Sessions

1. Defining a Rich Context Platform
   Session led by: Mark Grover, Lyft
   Room: Murrow
   - What would the ideal platform (akin to Amazon, Etsy, LinkedIn) do--both for the initial use cases, and also broadly useful across industry and academia.
   - What does an implementation look like? How readily can that be implemented?

2. Knowledge Graph
   Session led by: Daniel Vila Suero, Recognai
   Room: White
   - What is a knowledge graph?
   - Knowledge graph representation; embedding and inference; care and feeding.
3. Scholarly Infrastructure
   Session led by: Daniella Lowenberg, California Digital Library - University of California, Make Data Count, Dryad
   Room: Lisagor
   - How can metadata resulting from Rich Context augment the existing scholarly infrastructure?
   - Possible collaborations
4. Utility vs. Privacy
   Session led by: Brock Webb, OMB

16:00 -- Structured Sessions

1. Catalyzing a Community
   Session led by: Ed Kearns, Department of Commerce
   Room: Murrow
   - Provide for common needs in data infrastructure and metadata exchange.
   - Integrate across existing popular open source projects, in lieu of creating standalone projects.
   - Dependencies: JupyterLab, spaCy, PyTorch rdflib, etc. Standards: Egeria, W3C standards for metadata.
   - Frameworks: Amundsen and its emerging category. Approaches for evangelizing these projects together.
2. Centralized Services
   Session led by: Jan Höffler, ReplicationWiki
   Room: White
   - Identify where we need centralized services to augment what’s missing now, e.g., a global repository of datasets, with persistent identifiers.
   - Define a “service overlay” on the knowledge graph to complete the knowledge lifecycle.
3. Business Models
   Session led by: Josh Greenberg, Sloan Foundation
   Room: Lisagor

- Define business models that fit the problem and the community.
- Seeding initial work with research-funding support.
- How to become self-sustaining.

4. USDA Use Case
   session led by: Mark Denbaly, USDA and Patrick W. McLaughlin, USDA Economic Research Service

17:00 -- Wrap-Up Day 1

Saturday

09:00 -- Summary of Day 1 & Kickoff for Day 2

09:30 -- Lightning Talks (First Amendment Lounge)

1. Data Impact on Evidence-Based Policy
   Speakers: Stefan Bender, Deutsche Bundesbank & Jannick Blaschke, Deutsche Bundesbank
   Time: 09:30
    - Sound theory and adequate data are the backbones for evidence based policy making . In the research community, there is a well-established process to develop sound theory. But what about adequate data?
    - We present two applications bridging this gap, an empirical data impact factor and a research data finder (a data recommendation system for research).
2. Classifying Dataset Intent
   Speaker: Sebastian Kohlmeier, Allen Institute for Artificial Intelligence
   Time: 09:35
    - The objective of this talk is to highlight some initial exploratory work that we are doing at the Allen Institute for Artificial Intelligence to classify datasets based on their intent in a given research paper.
    - Presenting the classification categories that are being proposed along with their utility from a scientific research perspective.
3. All You Need is the Right Question
   Speaker: Haritz Puerto San Roman, KAIST
   Time: 09:40
    - Question-answer systems can be applied for text mining.
    - Presentation of how to utilize QA systems to retrieve datasets from publications.
4. Toward Direct Representation for Data and Data Sets
   Speaker: Bob Allen, NYU
   Time: 09:45

  - Techniques and standards for structured descriptions should greatly simplify indexing and improve access for data and data sets in text.
  - I will outline possibilities for going beyond traditional metadata and linked data to apply rich semantic "direct representation" to describe the way the data are collected and presented. Moreover, such knowledge structures can be supplemented with discourse and argumentation.

5. Building Google Dataset Search: Lessons Learned
   Speaker: Natasha Noy, Google
   Time: 09:50
   - Brief introduction of the Dataset Search and how it relies on the ecosystem of data providers to add metadata to the dataset landing pages.
   - Focus on the lessons learned, specifically on the fact that building such an ecosystem is much more of a social challenge than an engineering one.
6. Scientist First - Simple Product Principles for User-Centric HITL
   Speaker: Holly Corbett, ResearchGate
   Time: 9:55
   - I'll showcase the key things we've learnt from some of our successes and mistakes on the product side when working to incorporate user feedback into our training data. I'll then speculate on what this might mean for the rich context competition. I hope this will help frame later discussion about how the output of the competition can be used in real products into the future.

10:30 -- Structured Sessions Identify strategic focus areas for investment and development going forward

11:30 -- Closing Session & Call to Arms

12:00 -- Disband (optional lunch afterwards at ANXO)

Attendees

Amy Brand, MIT Press

Bob Allen, Yonsei University

Brock Webb, OMB

Cheryl Eavey, NSF

Chi-Ren Shyu, UMC

Chris Gorgolewski, Google

Christian Herzog, Digital Science

Christian Hirsch, Deutsche Bundesbank

Christian Zimmerman, Federal Reserve Bank of St. Louis

Christine L. Borgman, UCLA Information Studies

Dan Mbanga, Amazon AWS

Daniel Vila Suero, Recognai

Daniella Lowenberg, California Digital Library - University of California

Danny Goroff, Sloan Foundation

Deborah McGuinness, RPI

Dewey Murdick, Georgetown

Duane Williams, Digital Science

Ed Kearns, Department of Commerce

Euan Adie, overton.io

Eugene Burger, NOAA PMEL

Filippos Ventirozos, University of Manchester

Giwon Hong, KAIST

Gregory Gordon, SSRN

Haishan Fu, World Bank

Haritz Puerto San Roman, KAIST

Hendrik Doll, Deutsche Bundesbank

Holly Corbett, ResearchGate

Ian Mulvany, SAGE Publishing

Jan Höffler, ReplicationWiki

Jannick Blaschke, Deutsche Bundesbank

Jason Rhody, SSRC

Jo McEntyre, EMBL-EBI, EuropePMC, Project Freya

John Bohannon, Primer AI

John Kaufhold, Deep Learning Analytics at General Dynamics Mission Systems

Jonas Almeida, NCI DCEG

Josh Greenberg, Sloan Foundation

Julia Lane, NYU

Klaus Tochtermann, German National Library of Economics

Kris Rowley, GSA

Laura Noren, Obsidian Security

Laurel Haak, ORCID

Lorena Barba, GWU, NumFOCUS

Manish Parashar, NSF

Margie Graves, Office of Federal CIO

Mark Denbaly, USDA

Mario Diwersy, Digital Science

Mark Grover, Lyft

Matt Burton, University of Pittsburgh

Nancy Lutz, NSF

Nancy Potok, OMB

Natasha Noy, Google

Ophir Frieder, Georgetown University

Paco Nathan, Derwen AI, NYU

Patrick W. McLaughlin, USDA Economic Research Service

Pedro Gonzalez-Fernandez, Library of Congress

Philips Prasetyo, Living Analytics Research Centre

Rayid Ghani, CMU

Robert Stojnic, Atlas ML

Sam Klein, MIT Underlay Project

Sara Winge, Independent

Saul Shanabrook, Quansight

Sebastian Kohlmeier, Allen Institute for Artificial Intelligence

Sophie Rand, NYU

Stefan Bender, Deutsche Bundesbank

Stuart Feldman, Schmidt Futures

Sue Marquez, Rockefeller Foundation

Tim Janssen, Department of Defense

Tim Sehn, Liquidata

Tuan Do, UCLA

Tyler Christensen, NOAA

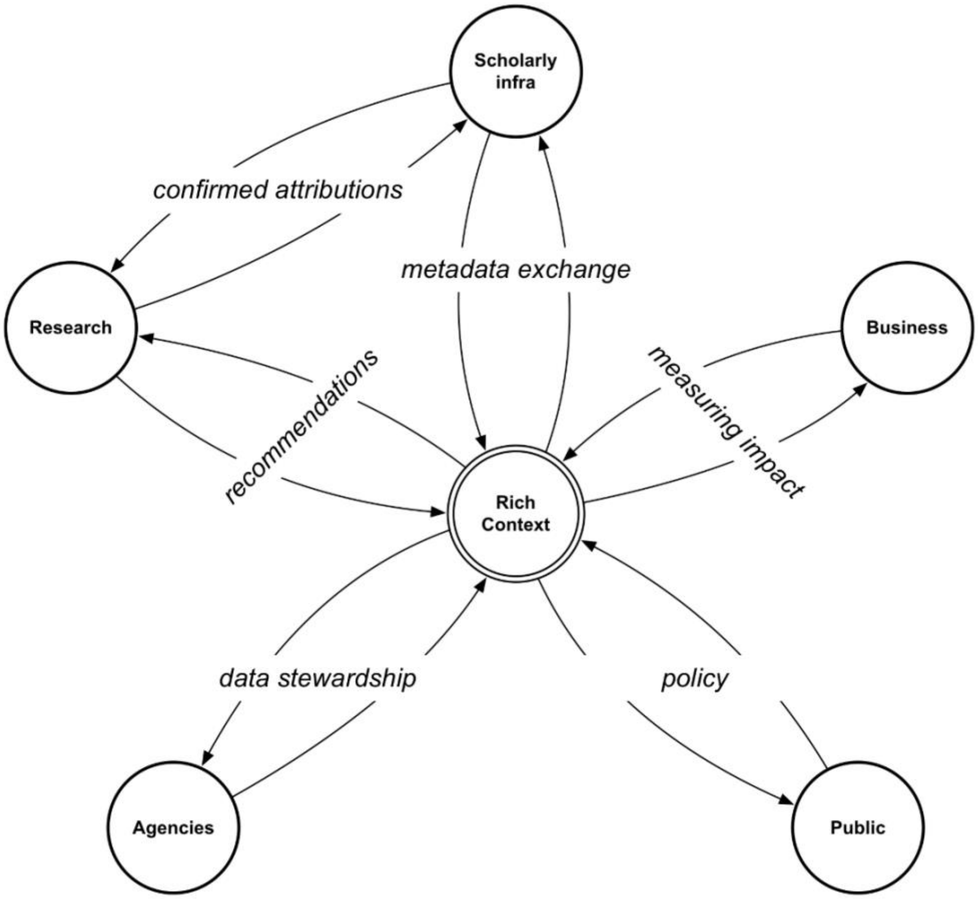
Waleed Ammar, Allen AI

Zach Sailer, Project Jupyter

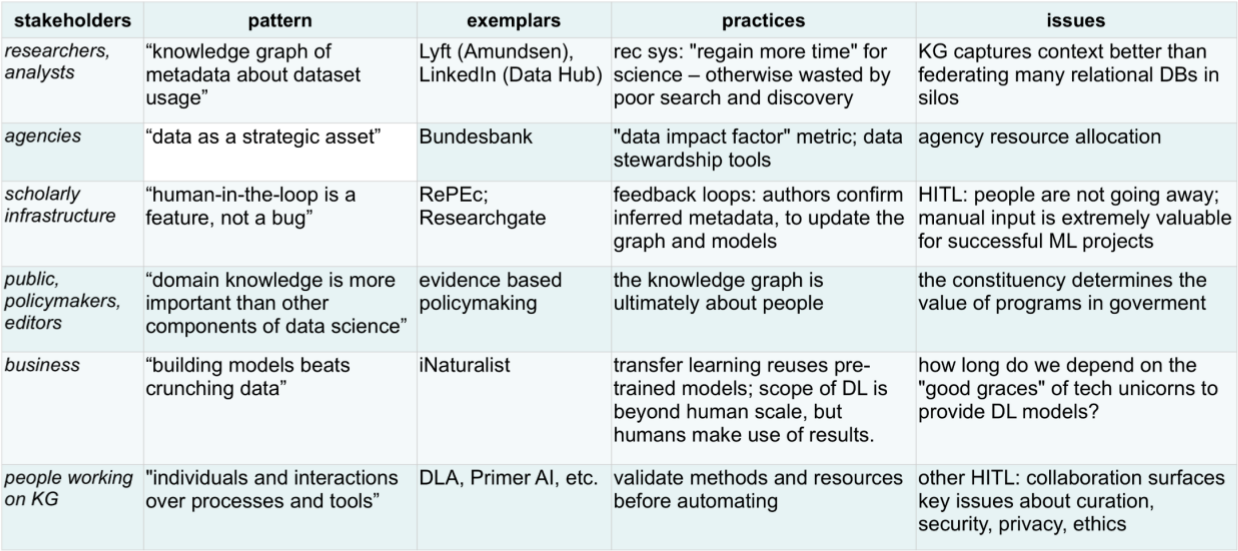
Akanksha Sharma, GSA

Visuals

Interactions

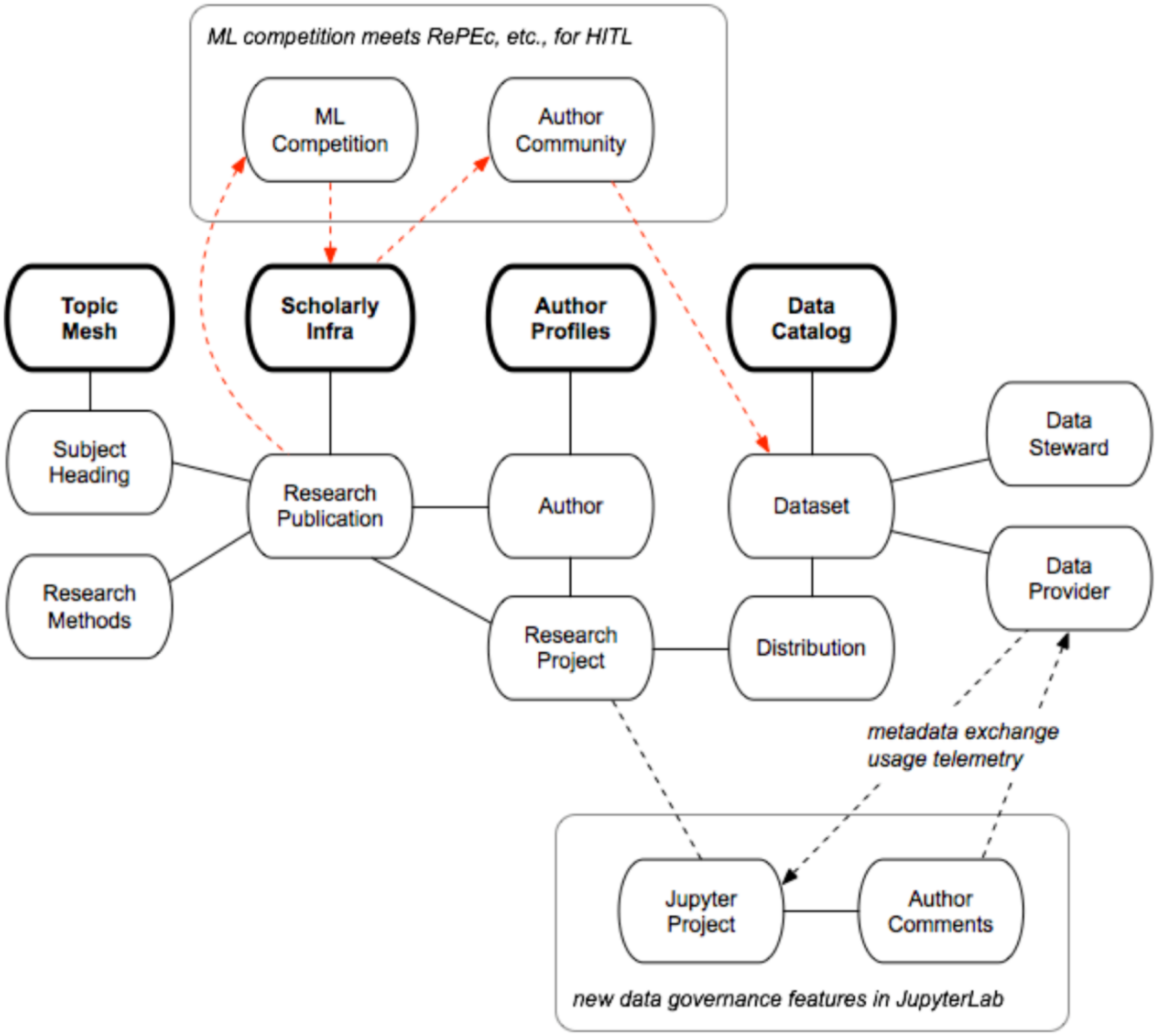


Affordances



stakeholders	pattern	exemplars	practices	issues
researchers, analysts	“knowledge graph of metadata about dataset usage”	Lyft (Amundsen), LinkedIn (Data Hub)	rec sys: "regain more time" for science – otherwise wasted by poor search and discovery	KG captures context better than federating many relational DBs in silos
agencies	“data as a strategic asset”	Bundesbank	"data impact factor" metric; data stewardship tools	agency resource allocation
scholarly infrastructure	“human-in-the-loop is a feature, not a bug”	RePEc; Researchgate	feedback loops: authors confirm inferred metadata, to update the graph and models	HITL: people are not going away; manual input is extremely valuable for successful ML projects
public, policymakers, editors	“domain knowledge is more important than other components of data science”	evidence based policymaking	the knowledge graph is ultimately about people	the constituency determines the value of programs in goverment
business	“building models beats crunching data”	iNaturalist	transfer learning reuses pre-trained models; scope of DL is beyond human scale, but humans make use of results.	how long do we depend on the "good graces" of tech unicorns to provide DL models?
people working on KG	"individuals and interactions over processes and tools”	DLA, Primer AI, etc.	validate methods and resources before automating	other HITL: collaboration surfaces key issues about curation, security, privacy, ethics

Knowledge Graph


ML competition meets RePEc, etc., for HITL
ML Competition
Author Community
Topic Mesh
Scholarly Infra
Author Profiles
Data Catalog
Subject Heading
Research Publication
Author
Dataset
Data Steward
Research Methods
Research Project
Distribution
Data Provider
metadata exchange
usage telemetry
Jupyter Project
Author Comments
new data governance features in JupyterLab