Rich Context in the Administrative Data Research Facility

ADRF Project Overview

The Administrative Data Research Facility (ADRF) provides a secure, cloud-based computing environment to analyze confidential micro-data for evidence-based policymaking, built atop Amazon GovCloud. All data access within the ADRF framework is based on the use of a *data stewardship* module which provides agency officials with an effective, streamlined workflow for authorizing access to sensitive data, monitoring appropriate uses of that data, then reporting on data usage in projects, while collecting the metadata generated by users.

Agencies need to be able to share data across state and jurisdictional lines in order to respond to many social problems. For example, examining the impact of access to jobs and neighborhood characteristics on the earnings and employment outcomes of ex-offenders and social benefit recipients on their subsequent recidivism or retention on welfare requires data from at least four different agencies (Corrections, Human Services, Labor and Housing) – ideally from multiple states. The same holds true for describing the earnings and employment outcomes of different education pathways, since students may get jobs in multiple states. To make that research possible, data from multiple organizations must be linked – with authorizations for data access based on requests at an individual, per-person level.

The design of the ADRF makes the necessary linkage of sensitive data possible, within a secure environment that maintains FedRAMP "moderate" certification. ADRF also streamlines the data sharing process, clarifying the relevant steps both for those who request data access and for those who determine whether to grant or reject that access. These functions are essential in that they provide necessary controls while also enabling straightforward answers to critical questions such as "Which projects use my data?" or "How is my data being used and which by products were generated by whom?"

To date, the ADRF platform has provided secure access for approved projects to over 50 confidential government datasets from over 20 different agencies at all levels of government.

Data Search and Discovery

Effective use of data depends on understanding how datasets have been produced and how they have been used in previous works. That understanding of data provenance is complicated by the fact that research often must link datasets from different data producers. Other factors confound this situation. On the one hand, the availability of inexpensive computing resources, ubiquitous connected mobile devices, social networks with global reach, etc., implies that researchers can acquire large, rich datasets. Researchers can also fit statistical models that might have seemed intractably complex merely a decade ago. On the other hand, accumulating important information about datasets, their provenance, and their usage has historically been a manual process. Sharing this kind of information across organizations is difficult in general, and when datasets include confidential data about human subjects it becomes impossible to provide open access to the original data. These issues combine to contribute to a lack of reproducibility and replicability in the study of human behavior, and threaten the legitimacy and utility of social science research.

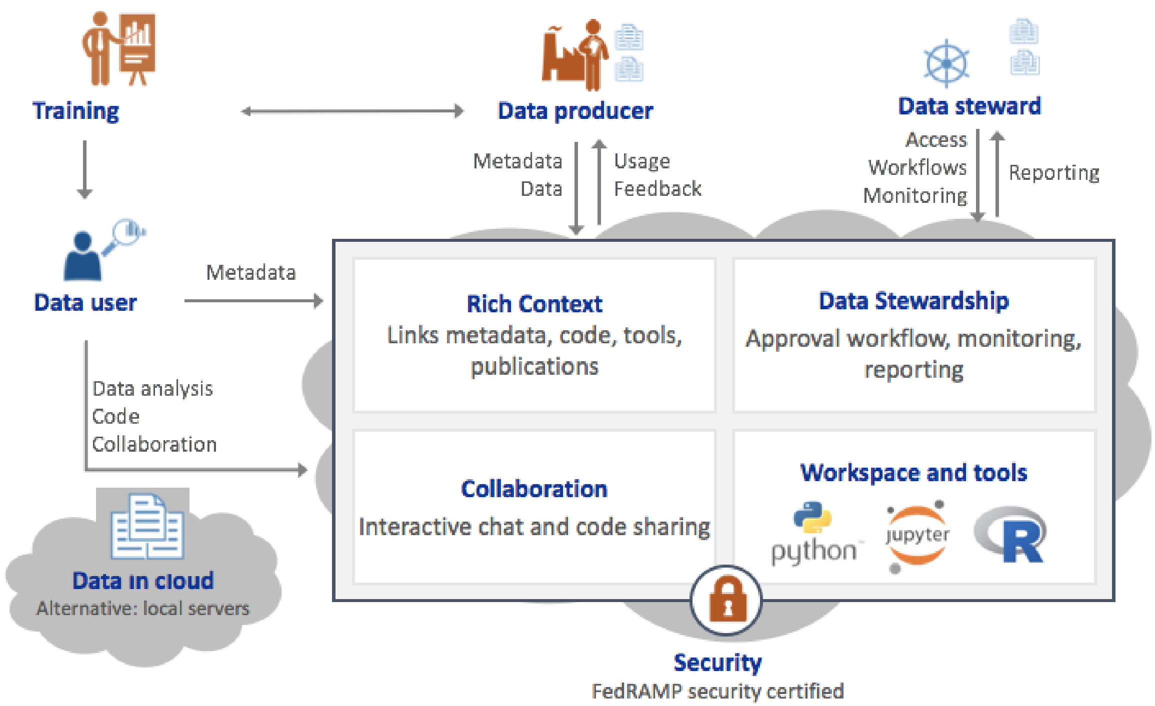
The core challenge is facilitating dataset search and discovery: the vast majority of data and research results cannot be easily discovered by other researchers. From one perspective, researchers are the users of micro data and its related *metadata* – in other words, information about the structure of datasets, their provenance, etc. – and those researchers produce outcomes, often in the form of publications. The information in empirical research publications is extremely valuable, because it includes information about which datasets are used, for which research topics, using which methods. These details produced through publications represent *metadata about datasets*. While the metadata within publications may be relatively unstructured – i.e., not explicitly articulated, nor shared outside of the current project – advances in machine learning provide means to extract metadata from unstructured sources.

The exchange of metadata plays another important role. From the perspective of a data publisher (i.e., an agency) the many concerns about security and data privacy mean that it is not possible for the general public to access sensitive data. Datasets which do not contain sensitive data may be made freely available to the public as *open data*. Sensitive datasets, however, have much more limited access, often with tiered levels of data availability. However, metadata for sensitive datasets may still be shared even though the data cannot be linked directly. So metadata provides a role of exchanging information about sensitive data, in ways that can be accumulated across a broader scope than individual research projects.

The opportunity at hand is to leverage machine learning advances to create feedback loops among the entities involved: such as researchers, datasets, data publishers, and publications. A new generation of tooling for search and discovery could leverage that to augment researchers: informing them about what datasets are being used, in which research fields, the tools involved, as well as the methods used and findings from the research.

The Rich Context approach

The workflows within the ADRF framework, illustrated in Figure 1, show how analysts and other researchers interact among themselves, data providers and data stewards.



Two components implement important point solutions for their respective requirements and users:

- **Collaboration and Workspace:** where researchers collaborate within a secured environment, having obtained authorizations via NDAs (non-disclosure agreements) or DUAs (data use authorizations).
- **Data Stewardship module:** where data stewards can review and determine whether to approve requests for using the datasets that they curate, and then monitor and report on subsequent usage.

These components of the ADRF framework represent *explicitly* linked feedback loops among the researchers, projects, datasets, and data stewards. Researchers also use other *implicitly* linked feedback loops externally to draw from published social science research. Overall, the general category of *linked data* describes these interactions.

A large body of AI applications leverages linked data. Related R&D efforts have focused mostly on public search engines, e-commerce platforms, and research in life sciences – while in social science research the use of this technology is relatively uncharted territory. Also, given the security and compliance requirements involved with sensitive data, the process of leveraging linked data in social science research takes on nuanced considerations and compels novel solutions.

We call this area of focus *Rich Context*: the interconnection of point solutions to facilitate research, as explicit feedback loops, along with means to leverage the implicit feedback loops that draw from published research. Making use of AI applications to augment social science research is the goal of Rich Context work, and that interconnection of feedback loops, through a graph, creates a kind of *virtuous cycle* for metadata – analogous to the famous virtuous cycle of data required for AI applications in industry, as described by Andrew Ng.

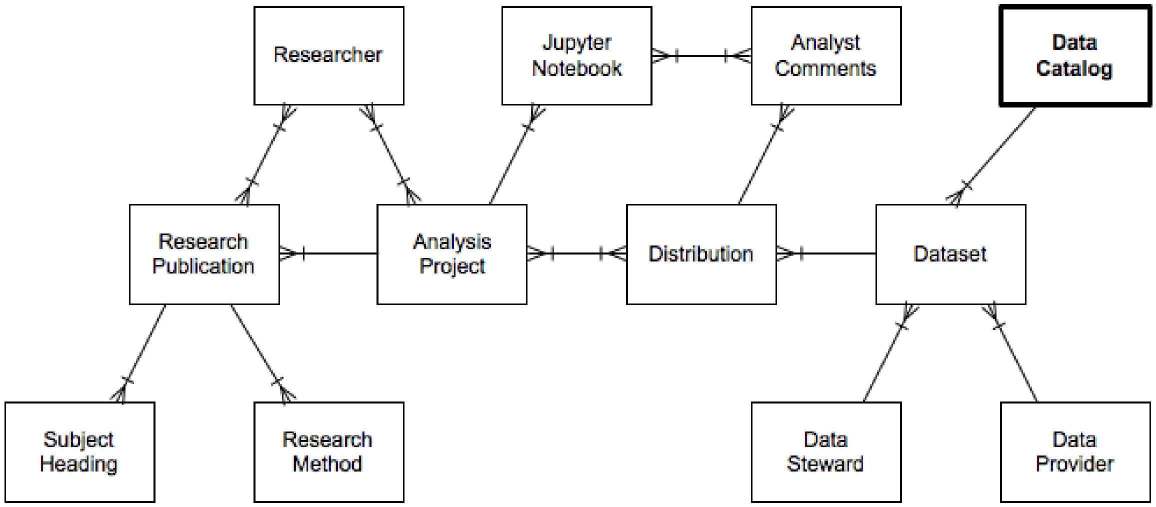
Rich Context as a Knowledge Graph practice

In general, guidance for Rich Context can be drawn from the FAIR[1] data principles for data management and data stewardship in science. The FAIR acronym stands for *Findable*, *Accessible*, *Interoperable*, and *Reusable* data, addressing the issue of reproducibility in scientific research. One observation from the original FAIR paper describes core tenets of Rich Context:

> *Humans, however, are not the only critical stakeholders in the milieu of scientific data. Similar problems are encountered by the applications and computational agents that we task to undertake data retrieval and analysis on our behalf. These 'computational stakeholders' are increasingly relevant, and demand as much, or more, attention as their importance grows. One of the grand challenges of data-intensive science, therefore, is to improve knowledge discovery through assisting both humans, and their computational agents, in the discovery of, access to, and integration and analysis of, task-appropriate scientific data and other scholarly digital objects.*

In other words, throughout the use cases for scientific data there are substantial opportunities for human-in-the-loop AI approaches, where the people involved increasingly have their work augmented by automated means, while the automation involved increasingly gets improved by incorporating human expertise. One can use the metaphor of a *graph* to represent the linkages: those that span across distinct research projects, those that require cross-agency collaboration with sensitive data, and those that integrate workflows beyond the scope of specific tools. Specifically, this work entails the development of a *knowledge graph* to represent metadata about datasets, researchers, projects, agencies, etc., – including the computational agents involved – as distinct entities connected through relations that model their linkage.

[1] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)



Much of the intelligence in this kind of system is based on leveraging inference across the graph, insights which could not be inferred within the scope of a single research project or through the use of one particular tool. Over time, the process accumulates a richer context of relations into that graph while clarifying and leveraging the feedback loops among the entities within the graph. Rich Context in the ADRF framework establishes foundations for that work in social science research.

The Rich Context Competition held during September 2018 through February 2019 invited AI research teams to compete in one aspect of Rich Context requirements. Several teams submitted solutions to automate the discovery of research datasets along with associated research methods and fields, as identified in social science research publications. Methods for machine learning and text analytics used by the four finalist teams provided complementary approaches, all focused on the problem of *entity linking*, with a corpus of social science research papers used as their training data.

The results of the competition provided metadata to describe links among datasets used in social science research. In other words, the outcome of the competition generated the basis for a moderately-sized knowledge graph. There are many publication sources to analyze, and the project will pursue that work as an ongoing process to extract the implied metadata. Meanwhile the increasing adoption and usage of the ADRF framework continues to accumulate metadata directly.

Use Cases for Rich Context

Looking at potential use cases for Rich Context more formally, we can identify needs for leveraging a knowledge graph about research datasets and related entities. For each of these needs, we can associate solutions based on open source software which have well-known use cases in industry.

As an example, consider a dataset *A001* published by a data provider *XYZ Agency* where *Jane Smith* works as the data steward responsible for curating that dataset. Over time, multiple research projects describe the use of *A001* in their published results. Some researchers note, on the one hand, that particular columns in data tables within *A001* have some troubling data quality issues – inconsistent names and acronyms, identifiers that require transformations before they can be used to join with other datasets, and so on. On the other hand, the body of research related to *A001* illustrates how it gets joined frequently with another dataset *B023* to support analysis using a particular research method. The two datasets provide more benefits when used together.

While access to the *A001* dataset gets managed through the *XYZ Agency* and its use of the ADRF framework, other datasets such as *B023* get used outside of that context. A knowledge graph is used to accumulate information about the datasets, research projects, the resulting published papers, etc., and applications for augmenting research derive quite directly from that graph. For example, feedback from researchers about how *A001* gets combined with other datasets outside of the *XYZ Agency* domain help guide *Jane Smith* to resolve some of the data quality issues. New columns get added with cleaner data for identifiers, which allows more effective linking. Other feedback based on machine learning models that have classified published papers then helps recommend research methods and candidate datasets to new analysts – and also to agencies that have adjacent needs, but did not previously have visibility into the datasets published by *XYZ Agency*.

These are the kinds of applications enabled through Rich Context. Search and discovery is clearly a need, although other use cases can help improve the discovery process and enhance social science research. The following sections discuss specific use cases and their high-level requirements for the associated technologies.

The Search and Discovery Vision

As described above, the vast majority of social science data and research results cannot be easily discovered by other researchers. While public search engines based on keyword search have been popularized by e-commerce platforms such as Google and Bing, the more general problem of search and discovery can be understood best as a graph problem, and the needs in social science research are more formally understood as recommendations across a graph.

For example, starting with a given dataset, who else has worked with that data? Which topics did they research? Which methods did they use? What were their results? In other words, starting from one entity in a knowledge graph, what other neighboring entities are linked?

These kinds of capabilities may be implemented simply by users traversing directly through the links of the graph. However, at scale, that volume of information can become tedious and overwhelming. It's generally more effective for user experience (UX) to have machine learning models summarize, then predict a set of the most likely paths through the graph from a particular starting point.

One good approach for this is the general case of *link prediction*[2]: given a researcher starting with a particular dataset and goals for topics or methods, represent that as a local, smaller graph. Then use link prediction to fill-in missing entities and relations, extending the local graph for that researcher. In other words, what other datasets should be joined, how can particular fields be used, what research topics or methods are related, which published papers might become foundations for this work? The most likely links inferred become top recommendations. Also, this kind of recommendation is not limited to the start of projects, it can be leveraged at almost any stage of research.

Entity Linking

The Rich Context Competition demonstrated how entities and relations used to construct a knowledge graph can be mined from a corpus of scientific papers. Machine learning methods for *entity linking*[3] used in the competition need to be generalized and extended, then used to analyze the ongoing stream of published social science research. This work provides potential benefits for the publishers, for example helping them analyze and annotate newly published papers, developing dashboards about data impact metrics for journals or authors, and so on.

An additional benefit of entity linking is to help correct abbreviations, localized acronyms, or mistakes in linked data references. This is an iterative process which will need integration and feedback with the Data Stewardship module in the ADRF framework.

Classifiers

As any researcher or librarian knows well, curating a large set of research papers by hand is labor-intensive and prone to errors. Machine learning models based on *supervised learning* or *semi-supervised learning* (human-in-the-loop) can produce classifiers that annotate research papers automatically.

At some point, the ADRF framework may run classifiers on the workflows (e.g., Jupyter notebooks) for projects in progress. By extension, classifiers may infer across the knowledge graph to add annotations for datasets as well.

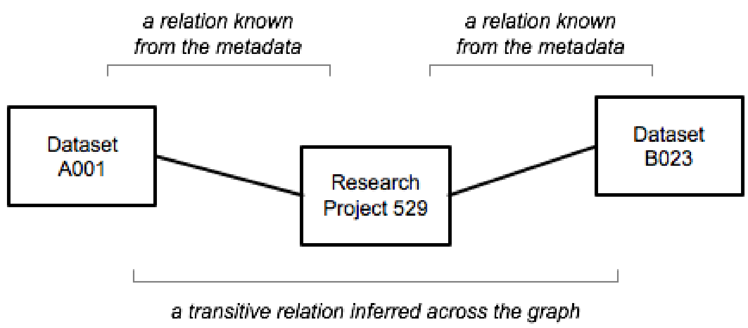
---

[2] For a sample of recent research papers regarding link prediction through graph embedding, see these Arxiv results.
[3] One of the better resources online for entity linking is NLP-progress which specifically tracks the state-of-the-art (SOTA) papers, along with their scores on recognized benchmarks.

This work can be considered a subset of link prediction, also related to entity linking.

Transitive Inference

The metadata collected through the use of the ADRF framework or extracted from research publications include relations that link entities in the graph. Once a graph is constructed, additional relations may be inferred. This is a case of *transitive inference*, which can help add useful annotations to the graph, as shown in the following diagram:



In an example from Norse mythology, Torunn is the daughter of Thor, and Thor is the son of Gaea, therefore Torun is the *granddaughter* of Gaea. The same process can apply, for example, to relations that describe links between datasets and researchers.

Note that embeddings have proven to be a powerful approach for inference about patterns, based on deep learning. On the current forefront of AI research, methods that leverage *reinforcement learning*[4] are positioned to outperform embeddings soon, since they explore/exploit the graph structure instead of relying on a history of observed patterns. This is especially useful for *knowledge graph completion*, where there are cases of incomplete metadata in the knowledge graph, which is essential for Rich Context work.

[4] For comparing between approaches based on RL vs. embedding, see "Multi-Hop Knowledge Graph Reasoning with Reward Shaping"; Xi Victoria Lin, Richard Socher, Caiming Xiong; *EMNLP 2018* arXiv:1808.10568 [cs.AI]

Iterative Improvement of the Knowledge Graph

Most of the finalist teams in the Rich Context Competition made use of other existing graphs to bootstrap their machine learning development work, such as the Microsoft Academic Graph, Semantic Scholar, and others purpose-built for the competition. Those teams cited how some graph would need to be extended in the future, to improve recognition accuracy.

Rich Context now subsumes that effort, making the iterative improvement of the knowledge graph an ongoing priority. In lieu of those other graphs used for bootstrap purposes during the competition, the Rich Context knowledge graph provides the foundation for machine learning.

This process of accreting more entities into the graph and refining their relations leads to better training data and improved machine learning models. Over time, as our models improve, the previously analyzed research papers can be re-evaluated to extract richer results. That work in turn enhances social science research within the ADRF framework, along with data curation. That overall dynamic represents the virtuous cycle of metadata, which continually improves Rich Context.

Axioms for Dataset Curation

Another immediate use of Rich Context is to assist the data stewards to understand the broader scope of usage for the datasets that they curate. For example, ontology *axioms* used on the metadata in the graph can help analyze:

- consistency checks for the incoming metadata
- which data stewardship rules apply in a given case

In a way, that helps codify what would otherwise be "institutional lore" – now captured for others to leverage, use in training new staff, etc.

Note that the ADRF framework must provide means for customizing and configuring these kinds of axioms, so that data stewardship rules rules are not tightly coupled with the ADRF code audit and release cycle. Those rules can change rapidly, depending on new legislation or other policy updates, or even due to different agency environments.

Leveraging Open Standards and Open Source

Overall, the Rich Context portion of the ARDF framework represents a *data catalog* along with associated *data governance* practices. As a first step in knowledge graph work, we can make use of existing open standards for metadata about data catalogs and datasets. For example, the W3C Data Activity coordinates a wide range of metadata standards, including:

- DCAT – metadata about data catalogs

- VoID – metadata about datasets
- DCMI – Dublin Core metadata terms
- SKOS – "simple knowledge organization system"

These represent controlled vocabularies described in OWL and based atop RDF. These standards can be combined and extended to suit the needs of specific use cases, such as within the ADRF framework. In particular, the Rich Context knowledge graph is a superset of a *DCAT-compliant data catalog*. Taken together, the localized extensions of these open standards represent an ontology – essentially as a specification for defining metadata that can be added into the knowledge graph and how that graph should be structured. Development of that ontology along with example metadata plus Python code to validate the graph is managed in the public repository adrf-onto on GitHub.

The workflows within the ADRF framework represent use cases of data governance, and there is substantial overlap between Rich Context and emerging trends for data governance in industry. There are open source projects which leverage knowledge graphs to collect metadata about datasets and their usage, where machine learning helps address the complexities[5] of data governance in industry data science work. For instance:

- Amundsen from Lyft
- Marquez from WeWork
- WhereHows from LinkedIn
- Databook from Uber (pending release for open source)

Of course the Rich Context work addresses special considerations for sensitive data and compliance requirements. Even so, much can be learned from these related open source projects in industry, which are pursuing similar kinds of use cases. TopQuadrant and AstraZeneca are examples of commercial vendors which construct knowledge graphs about datasets, also for data governance purposes – respectively in the Finance and Pharma business verticals. These commercial solutions similarly make use of DCAT, VoID, DMCI, SKOS, and also the FAIR data principles.

In general, the subject of metadata exchange for data governance use cases is addressed by the ODPi open standard Egeria and related work by Mandy Chessell[6], et al., including the Apache Atlas open source project. Much of that work focuses on standards used to validate the exchange of metadata reliably across different frameworks. This implies potential opportunities for the

---

[5] A good survey paper about these issues is given Ground: A Data Context Service, Hellerstein et al., *CIDR 2017*, based on research by UC Berkeley RISElab.

[6] See "The Case for Open Metadata", Mandy Chessell, *Frontiers in Data Science*, 2016-09-15.

ADRF framework to interoperate with other data governance solutions or related metadata services.

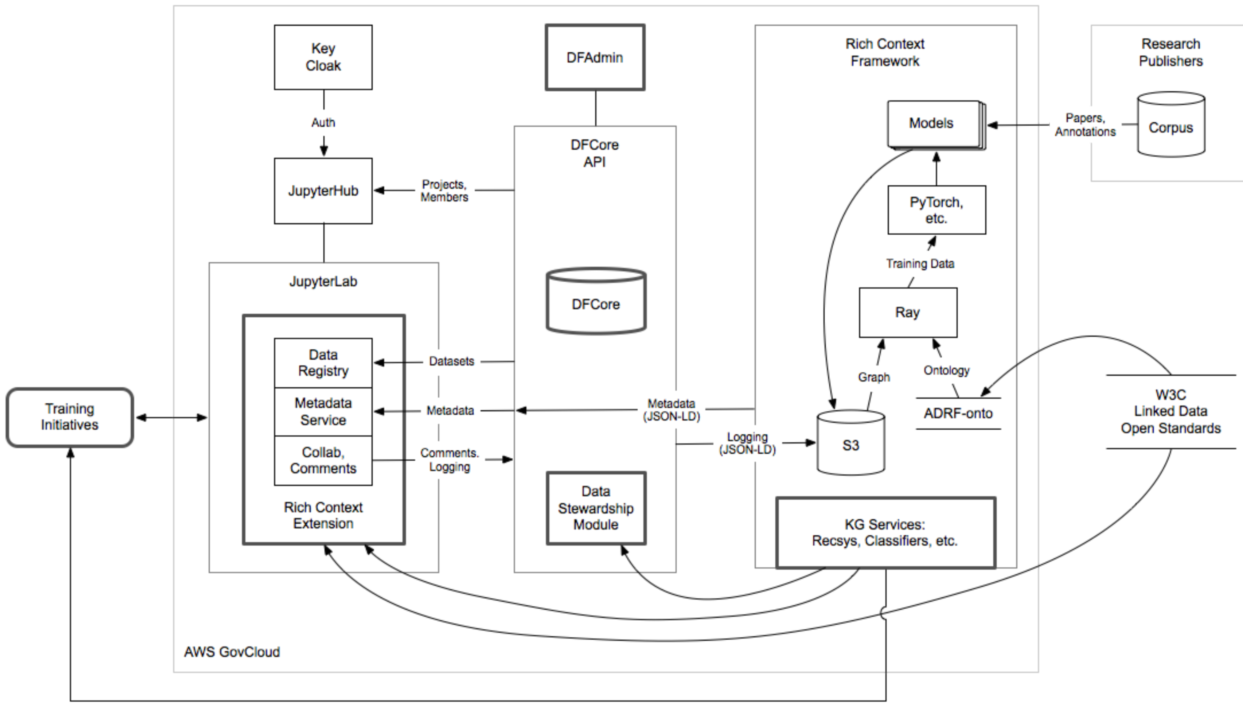
To help establish open standards and open source implementations related to Rich Context, the ADRF team has collaborated with Project Jupyter. A new Rich Context feature set is being added to JupyterLab, which is one of the key open source projects used in the architecture of the ADRF framework, and these new features will be integrated into its future releases. The new Rich Context features support projects as top-level entities, real-time collaboration and commenting, data registry, metadata handling, annotations, and usage tracking – as described in the Project Jupyter "press release" requests for comments: data explorer, metadata explorer, and commenting. For example, social science researchers working within the ADRF framework could use the commenting feature to make an annotation about data quality issues encountered in a particular dataset. That comment, as metadata about the dataset, would get imported into the knowledge graph, and could later be used for recommendations to a data steward or other researchers.

Note that most of the machine learning approaches referenced above are specific cases of *deep learning*, based on layered structures of artificial neural networks. In particular, *graph embedding*[7] is an approach that vectorizes portions of graphs to use as training data for deep learning models. Graph embedding can be used to perform entity linking, link prediction, etc. In many of these cases, the resulting machine learning models become proxies for the graph data, such that the entire knowledge graph data is not required in production use cases. That practice contrasts earlier and generally less effective approaches which relied on graph queries applied to the full data. Note that the winning team in the Rich Context Competition was from Allen AI which is a leader in the field of using embedded models for natural language. Typical open source frameworks which are popular for deep learning research include PyTorch (from Facebook) and the more recent Ray (from UC Berkeley RISElab).

System Architecture Overview

The following diagram illustrates a proposed system architecture for Rich Context as an additional module in the ADRF framework:

[7] For an overview of graph embedding, see "Graph Embedding for Deep Learning", Flawson Tong (2019-05-06).



Building on the DFCore features plus the Data Stewardship module, Rich Context provides both a destination for metadata (logging events from components, or extracted metadata from analysis of publications) and a source for metadata ontology used in the ADRF framework. Machine learning models get trained and updated based on the knowledge graph, then used for services (recommender system, classifiers, etc.) provided back into the ADRF framework, and additionally to support training initiatives – or for general purpose search and discovery by researchers.

The additional system components for implementing Rich Context are based primarily on open source software (e.g., PyTorch) and extensions of open standards (e.g., W3C), all within the security context of AWS GovCloud implementation of the ADRF framework.

Summary

Rich Context recognizes that social science research depends on *linked data* usage of micro data and its metadata. Effective management of that metadata is based on a graph that exists outside the context of component point solutions and specific workflows. While there is substantial use of linked data for ecommerce platforms and research in life sciences, social science research presents nuances and new challenges that haven't been addressed previously.

The Rich Context portions of the ADRF framework interconnect workflows that facilitate research – as explicit feedback loops in the graph – along with means to extract metadata from published research – as implicit feedback loops in the graph. That process creates a kind of virtuous cycle for metadata, making use of AI applications to augment social science research, with continual improvement of the entities and relations represented within the graph.

The first step was to create a corpus of research publications, used for training data during the Rich Context Competition, which demonstrated how to extract metadata from research publications.

The next step will be a formal implementation of the knowledge graph, based primarily on extensions of open standards and use of open source software. That graph is represented as an extension of a DCAT-compliant data catalog. It will eventually incorporate the new Rich Context features going into Project Jupyter. Immediate goals are to augment search and discovery in social science research, plus additional use cases that help improve the knowledge graph and augment research through the ADRF framework.

In the longer term, the process introduces human-in-the-loop AI into data curation, ultimately to reward researchers and data stewards whose work contributes additional information into the system. With this latter step, in the broader sense Rich Context helps establish a community focused on contributing code plus knowledge into the research process.