AMP: An Automated Metadata Pipeline

Beth Huffer [1] and Simon Handley [1]

[1] *Lingua Logica LLC, Denver, Colorado, United States*

Abstract

Making data more FAIR (Findable, Accessible, Interoperable, and Reusable) is key to helping data consumers make use of NASA data. The FAIR doctrine embraces the principle that facilitating machine-driven research activities is critical to supporting scientific research in the 21st century. Our automated metadata pipeline, AMP, generates syntactically and semantically consistent metadata records for U.S. National Aeronautics and Space Administration (NASA) Earth science datasets using ontologies and machine learning techniques. AMP addresses issues of usability and scalability for data providers and metadata curators who are asked to create robust metadata records to describe their data products, but who find it difficult to do so because of the lack of available tools. AMP auto-generates information-rich, semantically consistent metadata records for NASA datasets by sending the data through a semantic annotation pipeline that uses ontologies and machine learning techniques to generate sets of RDF assertions that describe it in detail. We demonstrate an end-to-end metadata curator's workflow, the final metadata records produced by AMP, and a data discovery and access application that makes use of those records.

Keywords

Ontologies, Machine Learning, Semantic Interoperability, Data Discovery, Earth Science, FAIR Data

1. Introduction

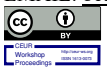
Our automated metadata pipeline, AMP, generates syntactically and semantically consistent metadata records for U.S. National Aeronautics and Space Administration (NASA) Earth science datasets using ontologies and machine learning techniques. Using AMP as it is deployed on Amazon Web Services (AWS), we demonstrate an end-to-end metadata curator's workflow, the final metadata records produced by AMP, and a data discovery and access application that makes use of those records. We will include interludes that provide insights into the back-end systems that make the production pipeline possible.

2. Background

NASA is a champion of free and open access to scientific data. Among the objectives identified in NASA's 2018 Strategic Plan are: safeguarding and improving life on Earth; and providing data and applications for operational use across a diverse set of communities of practice [1]. Achieving NASA's science goals - whether it be improving our ability to predict climate

FOIS 2021 Demonstrations, held at FOIS 2021 - 12th International Conference on Formal Ontology in Information Systems, September 13-17, 2021, Bolzano, Italy

EMAIL: beth@lingualogica.net (B. Huffer); simon@lingualogica.net (S. Handley)



change, or using Earth system science research to inform climate-related policy and planning decisions - depends on researchers, climate modelers, policy makers, environmental. planners, and others, making optimal use of the full range of Earth science data that NASA has to offer.

Making data more FAIR (Findable, Accessible, Interoperable, and Reusable) is key to helping data consumers make use of NASA data. The FAIR doctrine embraces the principle that facilitating machine-driven research activities is critical to supporting scientific research in the 21st century. Yet, the Earth science community still struggles to realize the FAIR objectives: scientific data discovery services are still inadequate, and scientists continue to spend a significant percentage of their time finding and preparing data for use in their research. In a study of science data users, Gregory et al. [2] found that

> researchers and environmental policy and decision makers need information from different locations and time, but they have difficulty accessing the information, or finding the right type [of information]… Integrating diverse data is problematic across the environmental sciences. Data collected at different scales and using different nomenclatures are difficult to merge (Dow et al., 2015; Maier, et al., 2014; Bowker, 2000b).

Implementing effective data discovery services and fully automated, machine-driven transactions starts with creating metadata that provides the information that users - both humans and machines - need to understand what is in a given dataset, and how to correctly use the data. But such metadata records are uncommon. More commonly, metadata records are inadequately contextualized, incomplete, or simply do not exist. Metadata requirements at data centers tend to be minimal, to ease the burden on data producers and metadata curators; and the metadata that does exist lacks adequate semantic underpinnings. As a result, tools and services for discovering and using data are, at best, syntactically interoperable; but they lack the semantic understanding necessary to achieve FAIR objectives. Unless the problem of semantic interoperability is addressed, scientific data discovery tools will continue to struggle to provide relevant search results, researchers will continue to struggle to make their data interoperate with their analytical tools, and NASA's goal of optimizing use of the full range of Earth science data that it has to offer will remain unrealized.

3. Automating the Metadata Production Process

3.1. Overview

Contributing to the problem of inadequate metadata is the fact that tools for generating metadata rely largely on manual curation and have little or no shared semantics. In some cases, metadata curators may be asked to pick from a controlled list of keywords; but the approach does not scale, and consistency is difficult to enforce. NASA archives currently hold over 8,000 data collections, many of which contain 100 or more individual datasets needing metadata. Manual curation of metadata for NASA datasets is not feasible. If metadata are to provide the means to address scientific data discovery and interoperability challenges at NASA, then scalable, user-friendly tools that can generate FAIR-compliant metadata are needed.

Our Automated Metadata Pipeline (AMP) addresses issues of usability and scalability for data producers and metadata curators who are asked to create robust metadata records to describe their data products, but who find it difficult to do so because of the lack of available tools. AMP auto-generates information-rich, semantically consistent metadata records for NASA datasets by sending the data through a semantic annotation pipeline that uses ontologies and

machine learning techniques to generate sets of RDF assertions that describe it in detail. To annotate the datasets in NASA collections, we look to the data itself to tell us what it represents. Our metadata production pipeline retrieves, parses, and subsets the datasets in large NASA Earth science data collections, and stages them in AWS S3 containers, where our machine learning (ML) component accesses and analyzes them. The ML component has been trained on a set of well understood, well characterized, and carefully chosen datasets that we call the "gold standard" datasets. A model of the temporal and spatial patterns in the data is derived from the gold standard datasets, and using this model, the ML component determines, for each new dataset (the "de novo dataset"), which of the gold standards the temporal-spatial patterns of the de novo dataset is most similar to. Given enough confidence in this similarity estimate, inference rules in the AMP Ontology attribute to the de novo dataset the same set of RDF assertions that describe the gold standard dataset. This set of assertions provides sufficient detail about the dataset to support highly precise data discovery, and to enable a downstream system to use the data. See Figure 1.

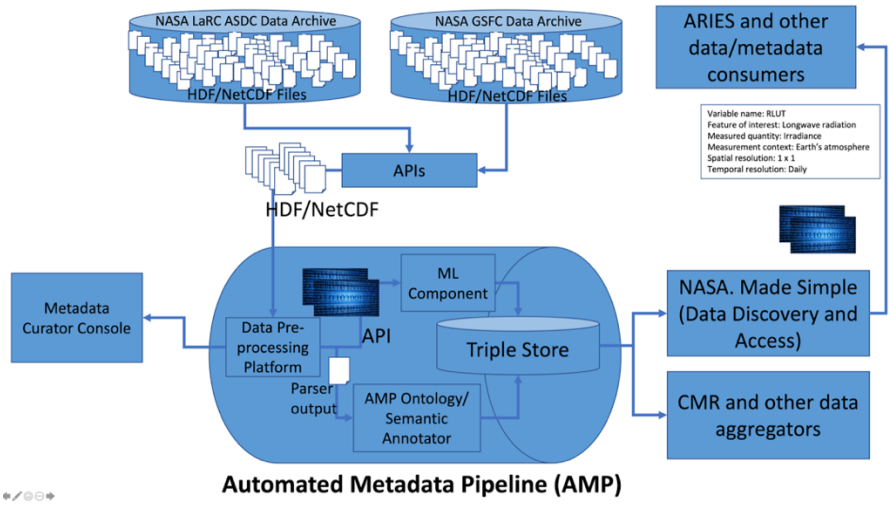


Figure 1. The AMP Workflow

3.2. Automating the Metadata Production Pipeline

Much of NASA's Earth science data is packaged in large collections, which assemble together numerous individual datasets for storage using an opaque format called HDF (Hierarchical Data Format). HDF is well-suited for cold-storage archiving, but is poorly suited for data analysis. To overcome this, we developed a sophisticated, serverless back-end system that retrieves and prepares the data for analysis by the ML component. The back-end system connects to NASA's 12 Distributed Active Archive Centers (DAACs) via Application Programming Interfaces (APIs), using the AWS API Gateway to accept incoming requests from a web-based metadata curator console. It executes a complex workflow that 1) populates a queuing system with file download URLs, 2) auto-scales AWS Fargate containers to download each file, 3) parses each file to extract the individual datasets (called "slicing"), 4) extracts information needed by the AMP ontology to generate inference rules, and 4) populates S3 buckets with the sliced datasets so they can be accessed by the ML component. A series of scripts convert the information extracted by the back-end, and the output of the ML component, into inference rules and assertions about the antecedent conditions that trigger the rules. These are pushed to the AMP ontology and used to generate the metadata records for each of the sliced datasets.

The AMP Ontology constitutes a conceptual model of the relationships between datasets and the Earth System observations that they record. Each dataset has an identifiable and well-defined type, with a unique set of membership criteria: i.e., a set of salient facts that determine, for any dataset and any dataset type, whether or not the dataset is of that type. Following the design pattern of the Semantic

Sensor Network Ontology [3], the salient facts about a dataset include, for example, the Earth System feature of interest or phenomenon that was measured by the sensor (or derived from a sensor measurement), e.g., water evaporation; the particular quantity of that feature that was measured (or derived) (e.g., mass flux rate); the medium or context in which it was measured (e.g., the Earth's Atmosphere); the vertical profile of the measurement (e.g., the Earth's surface); the process involved, if any (e.g., sublimation); and the spatial and temporal resolution of the data. Datasets that share those sets of facts are of the same type.

The ML component classifies a de novo dataset in terms of the set of gold-standard datasets: a de novo dataset *X* is predicted to have the same label (and therefore the same properties) as a gold standard *Y* if *X* is more similar to *Y* than to any of the other gold-standard datasets. Each dataset is a time-series of gridded measurement values that represent daily, monthly, 1-hourly, or 3-hourly averages, which can each be thought of as a temporal snapshot of the measured phenomenon. Each temporal snapshot in the gold standard datasets becomes a training example, and is given the same label as the dataset that it belongs to. All examples - gold standard and de novo - are spatially resampled to a 1.0°×1.0° grid, producing an input 180×360 matrix for datasets with global coverage. The classification of the de novo dataset is determined by decomposing it into a set of temporal snapshots, where each snapshot inherits its label from the dataset it belongs to. When training, each example is an input (the temporal snapshot, a 180×360 matrix) along with an output (the label or class). When doing inference, we compute a soft-max probability distribution for each temporal snapshot from the de novo dataset, combine these probability distributions into a distribution for the entire dataset, and use the per-dataset distribution to predict the class of the de novo dataset. The initial training uses 49 individual datasets from 8 NASA data collections which gives us 189,269 examples (labeled temporal snapshots), of which 50% were used for training, 25% for testing, and the remaining 25% as a hold-out set.

Once a de novo dataset has been classified with sufficient confidence, the ML component posts an assertion to the ontology via a SPARQL end-point, indicating which gold standard dataset the de novo dataset is most similar to. Upon receiving this information from the ML component, a set of inference rules, written as constructors in SPARQL Inference Notation, are executed to generate the appropriate assertions for the de novo variable. For example, this rule generates the set of assertions that indicate the feature of interest, measured quantity, and measurement context of the dataset:

```
CONSTRUCT { ?deNovo ?attribute ?value }
WHERE {
        ?deNovo AMP:matches ?goldStandard .
        ?attribute rdf:type AMP:AMPScienceProperty .
        ?goldStandard ?attribute ?value
}
```

Other inference rules assert properties specific to the datasets, in virtue of the collection they belong to. These rules are generated automatically using information extracted by the back-end pre-processing platform. Every data collection (i.e., a group of datasets packaged together in a set of HDF files) is created as a class in the ontology, of which the individual datasets are instances, and datasets in the collection share many common properties, such as spatial resolution. This rule, for example, about the spatial resolution of the data in the GPM_3CMB_DAY_V06 data collection is attached to an ontology class of the same name, so that each dataset that belongs to the collection will be annotated with an appropriate assertion about its spatial resolution:

```
CONSTRUCT {
    ?this AMP:spatialResolution AMP:Grid.25X.25Degree .
}
WHERE {
}
```

This rule about the instrument used to collect the data in the GPM_3CMB_DAY_V06 collection is also attached to it:

```
CONSTRUCT {
    ?this AMP:instrument AMP:GMIInstrument .
}
WHERE {
}
```

Using the combined techniques of the AMP end-to-end pipeline, we are able to auto-generate robust, detailed, and semantically consistent metadata records that drive a faceted search capability that enables users to specify precisely what kind of data they're looking for, and get back a set of results that actually satisfy their criteria. For example, a user can request measurements of black carbon mass concentration *in the atmosphere*, while excluding both organic carbon concentration *in soil*, and *optical depth* due to black carbon. This allows a researcher to quickly identify the *highly relevant* datasets, and even filter them further by specifying specific spatial resolutions, temporal resolutions, or instruments. We are implementing the faceted search capability in our prototype data discovery and access platform, which will be included in our demonstration.

AMP metadata records provide additional support for semantic interoperability by linking concepts referred to in AMP-generated metadata records to terms from well-established, external ontologies and glossaries such as the OBO Foundry's Environment Ontology (ENVO) [4], the Chemical Entities of Biological Interest (CHEBi) Ontology [5], the W3C Ontology for Quantity Kinds and Units [6], and the American Meteorological Society's Glossary of Meteorology [7]. These mappings help ensure that external systems already making use of those vocabularies can correctly interpret AMP-generated metadata. The URIs also provide additional information to metadata consumers about the definitions of the terms used in the metadata records, which we plan to make use of in future development to implement text search.

Our demonstration features an end-to-end run of the semantic annotation pipeline, with insights into the back-end mechanisms that make it work. We will also demonstrate our data discovery and access service, NASA Made Simple, which uses the metadata records we produce with AMP to drive a faceted search service that returns highly relevant results in response to user inputs.

4. Acknowledgements

This work was made possible through Grant No. 80NSSC20K0209 from NASA's Earth Science Technology Office Advanced Information Systems Technology Program.

5. References

[1] *NASA 2018 Strategic Plan*, 2018, URL: https://www.nasa.gov/sites/default/files/atoms/files/nasa_2018_strategic_plan.pdf
[2] Gregory, K. , Groth, P. , Cousijn, H. , Scharnhorst, A. and Wyatt, S. (2019), Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. Journal of the Association for Information Science and Technology, 70: 419-432. doi:10.1002/asi.24165
[3] Atkinson, Rob, García-Castro, Raul, Lieberman, Joshua, and Stadler, Claus, Semantic Sensor Network Ontology. URL: https://www.w3.org/TR/vocab-ssn/
[4] Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics, 4(1), 43. doi:10.1186/2041-1480-4-43
[5] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res., doi:10.1093/nar/gkv1031
[6] Lefort, Laurent, Ontology for Quantity Kinds and Units: units and quantities definitions. (2010), https://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu-rec20.html

[7] American Meteorology Society Glossary of Meteorological Terms, URL: https://glossary.ametsoc.org/wiki/Welcome