Improving Trust & Interoperability

Metadata for Data Refuge's Open Data Catalog

Sam Buechler and Joan Hua, University of Washington

Summary Report

Abstract

The Data Refuge Data Catalog archives federal climate and environmental data. It provides historical snapshots of datasets released on government data portals, which are vulnerable to deletion. Yet the metadata associated with these Data Refuge records have been minimal, and their relationship with the source records have not been clearly defined. To address this, we investigated crosswalking solutions, improved metadata of target datasets, customized an extensible schema, standardized tagging with controlled vocabulary, and documented workflow for future-phase implementation. The results improve trustworthiness and interoperability, facilitate more seamless data discovery and retrieval, and meet the needs of both archivists and researchers.

Process Overview

Data Refuge is a project that grew out of the Penn Program for Environmental Humanities (PPEH) at the University of Pennsylvania in Philadelphia. It was launched in 2016 due to growing concerns about losing full access to federal climate and environmental data as the presidential administration and priorities changed. The project began with a series of nationwide events called DataRescue,[1] about 50 of which had taken place by June 2017. These events included panel discussions and teach-ins about backing up public data locally, as well as crowdsourced efforts to crawl government data portals and seed the Internet Archive's End of Term (EOT) Harvest project. About 400 resulting datasets were uploaded to the Data Catalog.[2]

This Data Catalog holds historical snapshots of data from open data portals from federal agencies like the EPA and NOAA. The data objects need more descriptive information to further improve their usability, especially when and if the original datasets cease to be publicly accessible on government portals. While the previous workflow allows deeper-level URLs to be picked up and seeded to the EOT project, which solves a problem EOT had, the uploaded datas

[1] https://envirodatagov.org/datarescue/
[2] https://www.datarefuge.org/dataset

objects still lack metadata to describe their contents and to clarify relationships with the source records; they need more information beyond simply the original source and the URLs.

This Capstone[3] project is an effort to examine the metadata created through these crowdsourced DataRescue events and help the data achieve the goals of FAIR Data Principles[4] —findable, accessible, interoperable, and reusable. For the duration of this project, we focused on datasets harvested from the U.S. Environmental Protection Agency; the 66 data objects currently appear in the Data Catalog under said organization. We created tools (see **Toolset 1** and **Toolset 2** sections in this report) and workflows that can be applied in future phases to enhance the metadata of the entire Data Refuge Data Catalog.

With the FAIR Principle in mind, we also endeavored to prioritize trust, referencing considerations outlined by the CoreTrustSeal[5] trustworthy data repository requirements and ISO 16363:2012[6]. We determined the portion that we could improve for Data Refuge Data Catalog is the relationship between the harvested data and the equivalent datasets released on the United States open data portal (data.gov) and/or the Environmental Dataset Gateway (epg.epa.gov). These are not necessarily sources from which the data were harvested but corresponding datasets.

We investigated the relationship between the archived records in Data Refuge and the mirrored records on the source entity. We cross-referenced the two cata portals and gathered quivelant datasets to enhance the provenance documentation. By clearly defining this relationship, we can prepare the Catalog to be a more trustworthy data repository. The idea is that if records on federal data catalogs are lost—due to intentional or unintentional deletion—we have documentation of the lineage and relationships. This also allows us to compare the metadata on these portals.

We prioritized interoperability by choosing to create our metadata schema in a format that is platform-agnostic. This choice was then affirmed by the fact that CKAN may not be the permanent platform nor AWS the long-term hosting solution for the Data Refuge Data Catalog. The JSON schema format allows us to create metadata profiles independent of repository constraints. In this way we pave a path forward for machine-readability, metadata enhancement, and workflow integration with a new repository, such as the University of Pennsylvania institutional repository, in a systematic manner.

---

[3] The authors conducted this project, sponsored by Margaret Janz (Penn Libraries) and Carole Palmer (UW Information School), as their Capstone for their Master of Library and Information Science at the University of Washington. See the UW iSchool webpage about this project: https://ischool.uw.edu/capstone/projects/2020/improving-trust-interoperability-metadata-data-refuges-open-data-catalog.

[4] https://www.force11.org/group/fairgroup/fairprinciples

[5] https://www.coretrustseal.org/why-certification/requirements/

[6] https://www.iso.org/standard/56510.html

Observation 1: Existing Metadata

We determined that the metadata created by community volunteers have been bare-bones and not necessarily standardized. The metadata profiles consist of:

1. Descriptive metadata with six elements:
    a. TITLE [string]
    b. DESCRIPTION [string]
    c. TAGS [string] (community generated; include inconsistencies and misspellings)
    d. SOURCE [string with format: link]
    e. LAST UPDATED [date+time]
    f. CREATED [date+time]
2. Provenance metadata capturing the harvesting activity (with timestamp and harvester name) attached as a single JSON file to each data object for most, though not all, of the data packages
3. Additional metadata captured in the form of a PDF file attached to each data object for many of them; the PDFs were created by community volunteers when they interact with a mobile application—no longer available—designed specifically for the DataRescue events

We sought to expand on the required elements for descriptive metadata. The descriptions sometimes describe the tool or program from which the datasets were generated, while other descriptions focus centrally on the actual datasets harvested. In general, reading the description alone often does not provide enough insights to allow the user to understand what datasets exist in the package of the data object, what condition they are in, their size and format, and other important factors that helps user to compare the datasets and discern whether they should proceed to download the contents to answer their research question or information need.

In addition to having a more robust schema, we identified the need to standardize the tags to ensure they achieve the best usability. One term should be added to all objects for which the term is applicable, and we wanted to disambiguate related terms. Here are three examples of terms that are spread across the collection:

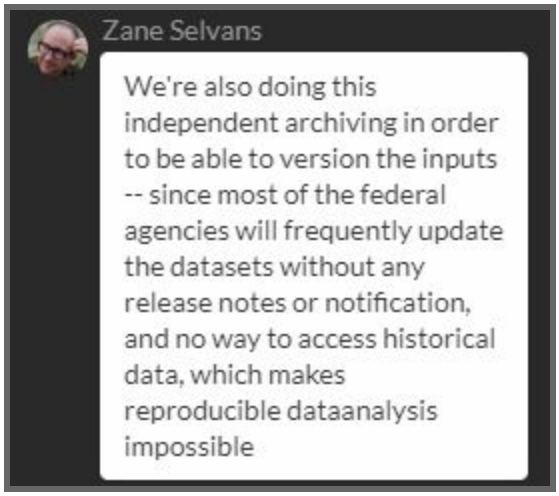
1. "water quality", "Water Quality", "water quality information"
2. "map", "mapping", "maps"
3. "carbon dioxide emissions"; "CO2 emissions"; "emission", "emissions", "emissions rates"

In the **Toolset 1: Subject Terms & Folksonomy** section below, we describe the hierarchical controlled vocabulary we created in more detail. We recommend using these for the **subject** property in the metadata schema created as part of this project. This would not require modification of the existing sets of tags, which we view as folksonomy. Retaining both records

allows us to honor the integrity of the Data Catalog, which bears strong characteristics of an archive of historical documentation.

Observation 2: Web Archive vs. Data Portal

Datasets released by the government on open data portals are vulnerable to deletion. The deletion can be unintentional or intentional, such as when a new administration's priorities change. This idea is what inspired the Data Refuge community efforts in the first place. In addition, the function of a government data portal is primarily to provide up-to-date information to allow for information sharing, transparency, and public scrutiny. It is, however, not specifically designed to be an archive of historical data. Data Refuge's Data Catalog fills this gap to create a safeguarded snapshot of the datasets available in 2017.



There is in fact a need for a place to access historical data. The image above shows a comment from Zane Selvans (Catalyst Cooperative)[7] during the *csv,conf* conference presentation "Getting climate advocates the data they need" on May 13, 2020. The comment illustrates this gap between a federal public data portal and a data archive that researchers can rely on over a long period of time.

There are examples of prior efforts that had goals similar to that of Data Refuge. Around the same time Data Refuge began its project, Climate Mirror[8] volunteers hosted downloaded federal data on their own servers. These data, however, lack provenance information to prove the datasets' validity and thus reusability by researchers.[9] A related repository is Datamirror.org, which was created in early 2017 at the University of California Curation Center (UC3) in collaboration with Code of Science & Society. It is described as an experiment, and the storage space, which housed 152,000 datasets totaling 42 TB, "has now reached the end of its service

[7] Catalyst Cooperative is a data science and policy consultancy focused on climate policies. Zane Selvans is their Chief Data Wrangler.
[8] https://climatemirror.org/
[9] Janz, M. (2018). Maintaining Access to Public Data: Lessons from Data Refuge. *Against the Grain*, *29*(6), 30.

life."[10] It is a much larger-scale project than our Capstone work. Whereas Datamirror.org's primary focus was data preservation, this Capstone project focuses on metadata quality.

Based on our research and discussions with project sponsors, we identified a unique hybrid characteristic of the Data Refuge Data Catalog: it is simultaneously viewed as a data catalog—similar to the government data portals or open repositories like Zenodo—and an archive of website snapshots and captures—similar to End of Term Web Archive[11] and perhaps webharvest.gov.[12]

Approaching our solutions required us to prioritize one purpose over another. We assessed the primary purpose and user base of the Data Refuge, especially in the future, beyond the 2017 harvesting communities. We determined that Data Refuge's Data Catalog now primarily functions as an archive of data and snapshots of data aggregators, tools, and website information from 2017. For researchers looking for datasets published by the government, they should still first go to the source entities, such as data.gov, to access them if they remain available there. This characterization is important to how we approached our solutions.

Response 1: Data Model

This hybrid nature led to some conundrums. For instance, when the collection includes a wide variety of data objects that range of structured datasets and harvested web content—such as EPA's staff list[13]—what authoritative terms can we use to distinguish between the various types? DataCite[14] terms for **resourceType** include *collection*, *interactiveResource*, *dataset*, and so on; DCMI terms for **type** similarly include *dataset* and *collection*; ISO 19115-1:2014 includes in its scope code values like *model* and *aggregate*. But we did not find a term that would accurately describe the capture of a web resource or aggregate—such as the Data Refuge version of TRI-CHIP[15]—that would effectively distinguish itself from the others that fit the type *dataset*.

---

[10] The Datamirror.org Experiment: Preservation Assurance for Federal Research Data. (n.d.). Retrieved May 23, 2020, from https://uc3.cdlib.org/2018/07/03/the-datamirror-org-experiment-preservation-assurance-for-federal-research-data/.

[11] The End of Term (EOT) Web Archive captures and saves U.S. Government websites at the end of presidential administrations (http://eotarchive.cdlib.org/). See how Data Refuge worked with EOT in M. Phillips & K. Phillips, 2018, End of Term 2016 Presidential Web Archive, *Against the Grain*, *29*(6), 27–30.

[12] Web Harvests (https://www.webharvest.gov/), initiated in 2004, is operated by National Archives and Records Administration (NARA).

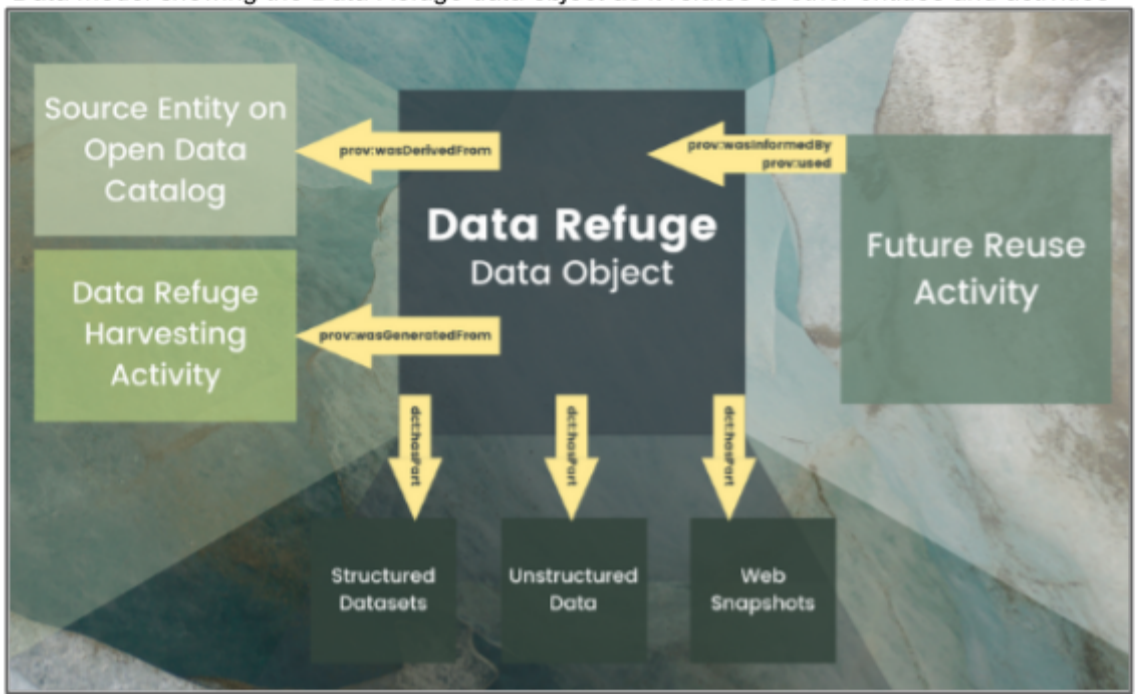
[13] Mohr, A. H. 2017. *Environmental Protection Agency (EPA) Staff Directory, 2017* (Data Refuge version). Retrieved May 31, 2020, from https://www.datarefuge.org/dataset/environmental-protection-agency-epa-staff-directory-2017.

[14] See DataCite Metadata Schema at https://schema.datacite.org/.

[15] Data Refuge admin3. 2017. *TRI-Chemical Hazard Information Profiles | Toxics Release Inventory Program* (Data Refuge version). Retrieved May 31, 2020, from https://www.datarefuge.org/dataset/tri-chemical-hazard-information-profiles-toxics-release-inventory-program.

With the Data Catalog's hybrid nature, we faced yet another conundrum: How should we parse the relationships between the files in each data package? How should we honor the integrity of the file structures of each data package, which is not consistent across the collection? Or does the structure suppress what would perhaps be the most important to retrieve, e.g., the structured datasets within each package? Are all files in each data package or object regarded as content and therefore data, or should certain files in the data object, such as the text files containing MD5 cryptographic hash values (sometimes automatically generated from a Data Conservancy Packaging Tool[16]) and the PDFs captured from the mobile application volunteers used to input information about each data object, be treated as contextual information and metadata?

Data model showing the Data Refuge data object as it relates to other entities and activities



The data model in the image above illustrates how we approach our work to design a metadata schema. In essence, we focus on describing and creating metadata for the data packages without privileging the specific structured dataset files in each package. Based on this data model, we view the datasets within each data object as components of the data object, as defined by the Dublin Core term **dct:hasPart**. In this model, all contents—whether they are the bagging information in text files or the structured datasets—are treated with the same level of importance. This is a different approach from another model to privilege the dataset—e.g., the "Structured Datasets" shown in this model—and consider the "Unstructured Data" and "Web Snapshots" as contextual information to define the "Structure Datasets." We choose the former

[16] https://github.com/DataConservancy/dcs-packaging-tool/releases

option based on our understanding of the important purpose of the Data Refuge Data Catalog as a type of archive of these data objects, as we outline in the above section.

We base this model heavily on the PROV data model (PROV-DM) for provenance. In particular, we borrow from the Derivations component[17] in order to specify the relationships between the data objects on Data Refuge that we seek to describe and their relationships with the entity from which the datasets were harvested and with equivalent or corresponding datasets currently on government data portals, such as data.gov and the Environmental Dataset Gateway. We propose using the predicate **wasDerivedFrom** to point to the entity that is the source record and **wasGeneratedFrom** to relate to the activities during the DataRescue events that harvested the datasets. This distinguishes a dataset or archived web resources on Data Refuge from a mirrored or even identical one on a live data portal. For example, Data Refuge's web capture of "Toxics Release Inventory (TRI) Program - Chemical Hazard Information Profiles (CHIP)"[18] **wasDerivedFrom** the TRI-CHIP program[19] that was (and still is at the time of writing) available on EPA's website. We further include a mechanism for recording information about equivalent datasets in the metadata schema we designed as part of this project, using the property **wasDerivedFrom** from the PROV XML schema.[20] We discuss this in further detail in the section **Toolset 2: Proposed Metadata Schema** below.

Response 2: Comparing Resources

At the start of this Capstone project, we recognized that most datasets harvested have not been deleted from government portals like data.gov and that data.gov happens to utilize the same platform, CKAN, as Data Refuge does. Considering these advantages, we compared both repositories to see how we might systematically map descriptive information from data.gov to Data Refuge and enhance the metadata in the Data Catalog.

The data objects on Data Refuge often point to two sources: One is a URL in the metadata that links to a web resource or tool pertinent to the data object; in the case for EPA datasets, this can still exist and remain updated on epa.gov. Another one is an Internet Archive URL that, if available, appears in the description field and points to historical views of the tool when it or the datasets aggregated through it was harvested. The values in the **source** field for each Data Refuge data object are often web pages or web resources, and they almost never connect directly to a dataset on data.gov. We manually searched for corresponding datasets on data.gov and edg.epa.gov and gathered links in order to compare the metadata profiles of similar

[17] See *PROV-DM, 5.2 Componenet 2: Derivations* at https://www.w3.org/TR/prov-dm/#component2.
[18] The same data package was referenced in note 15; see https://www.datarefuge.org/dataset/tri-chemical-hazard-information-profiles-toxics-release-inventory-program.
[19] *TRI-Chemical Hazard Information Profiles (TRI-CHIP)* (EPA version). Retrieved May 31, 2020, from https://www.epa.gov/toxics-release-inventory-tri-program/tri-chemical-hazard-information-profiles-tri-chip.
[20] http://www.w3.org/ns/prov#

datasets. It was often the case, however, that the datasets on data.gov or edg.epa.gov were related but not exact, and there was not a consistent pattern as to which data package might have the same mirroried records on the data portals (see table below for examples). Therefore, we concluded that comparing the data objects and determining whether the matdata are reusable on Data Refuge are tasks that still require heavy human intervention and cannot be automated at this stage.

Example comparisons of Data Refuge objects and other source entities

Data object	Sources provided on Data Refuge	Compare with EDG	Compare with data.gov
Toxics Release Inventory (TRI) Pollution Prevention Facilities	**Source:** https://www3.epa.gov/enviro/facts/tri/p2.html [link now redirects to Envirofacts] **Internet Archive link:** https://web.archive.org/web/*/https://www3.epa.gov/enviro/facts/tri/p2.html [no longer works]	[No equivalent]	[No equivalent]
Enforcement and Compliance History Online	**Source:** https://echo.epa.gov/ [the data package contains datasets downloaded to include years 2009–2017] **Internet Archive link:** https://web.archive.org/web/*/https://echo.epa.gov	https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B5DF97189-37B4-421E-B74C-C225EECC423F%7D [Equivalent]	https://catalog.data.gov/dataset?organization=epa-gov&q=%22Enforcement+and+Compliance+History+Online%22 [Search results listing reports from different years]
CPCat: Chemical and Product Categories	**Source:** https://actor.epa.gov/cpcat/faces/download.xhtml **Internet Archive link:** https://web.archive.org/web/*/https://actor.epa.gov/cpcat/faces/download.html [no longer works]	https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B90CDFB11-E94E-4E84-942C-5B3D2B5ED0CD%7D [Equivalent]	https://catalog.data.gov/dataset/consumer-product-category-database-8a556 [Equivalent]

It may be possible to isolate and target data objects that have direct equivalent records on data.gov, batch download their metadata files in JSON using API, and then batch update data objects on Data Refuge. Even though it is possible to aggregate the data.gov metadata in JSON Line files exposed by the CKAN API,[21] we believe this workflow currently would not be useful because there is not a single distinguishing attribute that we were able to identify that would isolate all datasets on data.gov that exist as a derivation on Data Refuge. The relevant datasets still require manual comparison and identification, and the 66 EPA datasets we sampled reveal that not many have direct equivalent records on data.gov that contain metadata we could directly apply to Data Refuge. Furthermore, we learned from our stakeholders that the CKAN platform Data Refuge currently uses is not a long-term solution. This led us to create the toolset (see sections below) that offer a more platform-agnostic path forward for enriching metadata for the Data Refuge Data Catalog.

[21] Read about Data Harvesting on data.gov documentation for developers here: https://www.data.gov/developers/harvesting.

Toolset 1: Subject Terms & Folksonomy

Because of the information derived from comparing resources, we developed a strategy for creating elements of descriptive metadata that both improved and relied on the existing information in the Data Catalog. Data Refuge, in its current iteration, makes use of a folksonomy tagging system. As described in **Observation 1**, this tagging system is incomplete and ambiguous in nature but it does provide some key archival information about the resource and its harvesting. Instead of removing the existing tags altogether, it was determined to create a supplemental controlled vocabulary to be applied under the proposed Data Refuge Schema **dc:subject**.

The controlled vocabulary, informally named the Acronym Controlled Vocabulary, was created to meet this supplemental need. In its current version (v02), the Acronym Controlled vocabulary assists in indexing Data Refuge Data Catalog objects to improve trust and interoperability of the catalog. Based on needs defined by Data Refuge stakeholders, the Acronym Controlled Vocabulary is composed solely of abbreviated terminology found in the existing descriptive metadata of the catalog. Given the scope of this project, the Acronym Controlled Vocabulary is only recommended for use on the data packages within the Data Catalog that were harvested from the U.S. Environmental Protection Agency.

Constructing the Acronym Controlled Vocabulary involved analysis of descriptive metadata and source links within the Data Refuge Data Catalog for any named agencies, organizations, programs, or information resources. Once these were identified, the terms collected were checked against EPA resources for validity, relevance, and potential disambiguation issues. The final step was confirming acronyms using the EPA System of Registries'[22] Terms & Acronyms resource.[23] For information that was found to be within the scope of the EPA, but not a sole EPA entity, USA.gov[24] was used to identify governing agencies and relevant resources accordingly.

More information about using, expanding, and revising the Acronym Controlled Vocabulary can be found in the **acronymControlledVocabulary_v02**[25] document.

Toolset 2: Proposed Metadata Schema

In the above sections of this report, we describe the need to expand upon the basic metadata that are provided with each data object in the Data Refuge Data Catalog. This would not only

[22] See *System of Registries* at https://ofmpub.epa.gov/sor_internet/registry/sysofreg/home/overview/home.do
[23] See *Terms & Acronyms* at https://ofmpub.epa.gov/sor_internet/registry/termreg/searchandretrieve/termsandacronyms/search.do
[24] See *USA.gov* at https://www.usa.gov/
[25] See file *acronymControlledVocabulary_v02* at https://github.com/jo-hua/dataRefugeCapstone

allow each record to provide more robust descriptive, technical, and contextualizing information, but it would also allow the objects in the collection to be more discoverable across the various organizations listed in the Catalog and across the repolisory. To enable this, standardization of the metadata—in terms of syntax, semantics, and application—is important. We also describe in this report the desire to systematize the workflow in order to scale up and facilitate future-phase application. Taking into considerations of standardization, machine-readability, feasibility (for a team of volunteers, students, and staff that may experience turnovers), and compatibility with future repository platform(s) that is yet to be confirmed, we propose a customized Data Refuge Schema written in the JSON format.[26] The schema is accessible on the GitHub repository mentioned in **Toolset 1** above and via our project sponsor at Penn Libraries/PPEH.

The proposed Data Refuge Schema contains 15 required properties: **@type** (type of metadata), **title**, **identifier**, **packageId**, **date**, **format**, **extent**, **dataType**, **description**, **subject**, **temporalCoverage**, **organization**, **wasDerivedFrom**, **collection**, and **license**. The document includes descriptions of the properties, instructions on how they should be applied, and, when appropriate, a default value. We endeavor to map metadata properties to existing standards and namespaces. As such, the Data Refuge Schema utilizes properties from the Project Open Data Metadata Schema v1.1, Dublin Core (DCMI), Data Catalog Vocabulary (DCAT), Ecological Metadata Language (EML), and the PROV-XML Schema. This mapping information is represented by **conformsTo** in the schema document.

The design process of this proposed schema involves evaluating the additional information stakeholders would find useful or find the current metadata to be lacking. For instance, in discussion with our sponsor at Penn Libraries/PPEH, we quickly identified file size—which varies widely in the collection—as a key piece of metadata that should be added to increase ease of use, as it would help determine whether or not the user could proceed to download the datasets. This became the property **dct:extent** that we include in the Data Refuge Schema.

We determined it would be important to require **eml:temporalCoverage** because of the fact that—as we explain in earlier sections of this report—the Data Refuge data objects are treated as archived, historical datasets and are not guaranteed to represent current data. The schema also takes into account the potential for the Data Refuge Data Catalog to become part of other broader repositories. The PROV properties are meant to record this information to be associated with each data object, regardless of whether they become part of a different collection structure.

Finally, the **dcat:dataType** property allows the type of object to be more accurately specified. In the **Observation 2: Web Archive vs. Data Portal** and **Response 1: Data Model** sections above, we discuss the unique hybrid nature of the Data Refuge Data Catalog and how that warrants distinguishing terms to describe a snapshot of a web resource or interactive database,

[26] The schema document follows JSON Schema 7.0 syntax. For more, see M. Droettboom, 2020, *Understanding JSON Schema*, JSON Schema. Retrieved May 31, 2020, from https://json-schema.org/understanding-json-schema/.

amid other structured datasets (for which the value *Dataset* is used for the **dcat:dataType** property). Despite searching and consultation with experts, we have yet to identify an authoritative term that is used this way. For now, we recommend using *Web capture* consistently as the term for this type of data object. An example of this application is shown in **datasetNum28.json**[27] on the same GitHub repository.

Partial view of a metadata file showing how the schema is used

As with the controlled vocabulary, this schema is meant to be revised as the Data Refuge team uses it to apply metadata to the rest of the collection. We recommend the team work through their documents and versioning on their official GitHub repository. We also provide an additional checkpoint to validate the schema document; see **ValidateJsonSchema.ipynb** in the repository and documentation of using jsonschema[28] (and implementation of JSON Schema for Python) to validate a new version. There are also various web-based JSON validation tools that can be utilized for this purpose.

Next Steps for Data Refuge & Future Considerations

Metadata Profiles

The logical next step to this project is the implementation of metadata profiles using the JSON schema and controlled vocabulary. We have provided examples of the JSON schema as

[27] https://github.com/jo-hua/dataRefugeCapstone/blob/master/datasetNum28.json
[28] Project page for jsonschema is at https://pypi.org/project/jsonschema/, which links to the GitHub repository; see also Schema Validation at https://python-jsonschema.readthedocs.io/en/stable/validate/.

applied to four data packages[29] within the catalog and the Acronym Controlled Vocabulary documentation provides detailed instruction for use in indexing. As it stands, the JSON schema is ready to be applied to all data packages within the Data Refuge Data Catalog.

If desired, it may be helpful to consider the creation of a workflow document for filling out the JSON schema. That being said, the original schema documentation[30] provides information on how to fill in the required fields and can serve as this documentation for those that are comfortable with reading JSON formats.

Further into the future, the schema and metadata files would require systematic validation to ensure that the files conform to both the JSON Schema specifications and the specifications of the proposed Data Refuge Metadata Schema. Continuous application of the metadata and considerations for how the files would be appended to the data objects to be used by viewers as well as serve administrative purposes would depend on the permanent platform and repository choice for the Data Refuge Data Catalog.

Controlled Vocabulary

Since the current version of the Acronym Controlled Vocabulary is only useful for data packages harvested from the EPA, it's important to expand it to include the other organizations[31] that data packages have been harvested from within the Data Refuge Data Catalog. Validation of this controlled vocabulary using a semantics platform (e.g., PoolParty[32]) will also be necessary. For validation resources, it's recommended to use the descriptions and source links as that is the domain of the vocabulary.

In previous considerations of a controlled vocabulary, it was discussed to potentially add topic-specific terms to better define the type of information contained within the data packages. Unfortunately this fell outside of our scope but through our explorations we have suggestions for future implementation if desired. Adding a Classification Research Group facet[33] pertaining to entities, things, and objects to the existing controlled vocabulary could be a functional option as seen in the Pacific Climate Change Portal Topics Controlled Vocabulary.[34]

---

[29] See *datasetNum12.json, datasetNum28.json, datasetNum35.json,* and *datasetNum41.json* at https://github.com/jo-hua/dataRefugeCapstone

[30] See *dataRefugeSchema_v1.0.json* at https://github.com/jo-hua/dataRefugeCapstone

[31] See *Data Catalog Organizations* at https://www.datarefuge.org/organization.

[32] See *PoolParty Semantic Suite* at https://www.poolparty.biz/.

[33] Aitchison, J., Gilchrist, A. & Bawden, D. 2000. Section F: Structure and relationships in *Thesaurus Construction and Use : a Practical Manual,* 4th ed. (p. 49-68) Chicago; London: Fitzroy Dearborn. - many resources contain information on CRG but this was the resource used when building the Acronym Controlled Vocabulary

[34] The *Pacific Climate Change Portal Topics Controlled Vocabulary* was created specifically for describing objects in the PCCP. See the *Topics Controlled Vocabulary* at https://www.pacificclimatechange.net/sites/default/files/documents/PCCP%20Topics%20Controlled%20Vocabulary%202.0.pdf

Access

For continual access to the resources, materials should be pulled from the supplied GitHub and linked to one owned by either Penn Libraries, Data Refuge, or PPEH. This will ensure perpetual access as well as application to organization specific versioning and maintenance procedures.