SEAD Sustainable Environment Actionable Data

Towards Universally Accessible Data Publications

The National DATA SERVICE

James D. Myers, Ph.D.
Interuniversity Consortium for Political and Social Research
University of Michigan
myersjd@umich.edu

Abstract

Many organizations including the National Data Service share a vision in which scientific data is easily citable and machine retrievable – enhancing research reproducibility, providing reusable assets for next-generation research, and accelerating scientific progress by facilitating the flow of data through an ecosystem of scientific services and applications. To further elucidate this vision, an analogy is often made with the benefits of digital publication of scientific papers, where the combination of citable persistent identifiers and discovery services are the primary keys to accessibility. However, data is different. Its scale, heterogeneity, and complexity will require that further steps in exploration, retrieval and interpretation be handled by machine. Unfortunately, today there is little standardization in data publication mechanisms or in the publication themselves that would enable more than point solutions for these further steps. The NDS Universally Accessible Data Publications Pilot project has identified several areas where standardization may be possible, across existing data repositories and services, that could simplify data publication interchange and motivate further work to harmonize data publication practices. This poster introduces the UADP effort in terms of motivation and goals and then uses work done within the SEAD DataNet project on a model and repository agnostic, standards-based approach to data publication to highlight the potential and limitations of some of the technologies and approaches currently being discussed in the UADP effort.

Universally Accessible Data Publications (UADP)

The UADP Pilot project was launched after discussions at the NDS Consortium's 6th Workshop. Initial participants have developed web materials to document the scope of the effort, reviewed relevant technologies and other standardization efforts, and begun discussions to drive toward consensus solutions. Additional participation is being actively recruited. See *https://nationaldataservice.atlassian.net/wiki/display/NDSC/Universally+Accessible+Data+Publications+Pilot* .

The Challenge: Given the persistent identifier of an arbitrary data publication (DOI, Handle, ARK, etc.) there is no standard way for the data and metadata it represents to be retrieved by computer.

Use Cases: Scheme-specific mechanisms exist to retrieve some discovery metadata and to find a human-readable 'Landing Page' for the publication, but the following use cases cannot be addressed today:

- A researcher drops the identifier for a data publication from an arbitrary source in their analysis tool and the tool is able to retrieve and process a relevant data file(s) automatically or after presenting the researcher with a brows-able display of the publications content so a selection can be made.
- The results from a data analysis such as the one just described can be published in a way that they can be retrieved and then analyzed or visualized within another service, duplicating the first use case, without any coordination between the service providers.
- A domain-specific catalog is able to discover all data publications in targeted repositories and perform a deep scan of their content to identify and index any relevant content.

UADP Pilot Project Goals:

- A standard mechanism, given a data publication identifier (in any scheme) to resolve the available metadata
- A standard means of harvesting available data publication identifiers from a given repository
- Documentation of the current practices of the repositories and services for producing and consuming metadata represented by the pilot participants and identification of existing best practices and areas of consensus
- Implementation, possibly across a subset of the group, of a standard means to discover and retrieve individual files within a data publication.

Sustainable Environment Actionable Data (SEAD)

The Sustainable Environment Actionable Data (SEAD) project, initiated through the NSF DataNet program, has developed a highly flexible approach to data publication that allows individual groups to customize the documentation and structure of their publications and to request publication through one of several repositories. The features and architecture of SEAD's services, which have evolved significantly between versions 1 and 2, have been described elsewhere (http://sead-data.net/about/publicationspresentations).

Most relevant to this discussion is SEAD's core data model which supports manual and automated annotation of data with arbitrary key/value metadata from user-specified vocabularies, and a basic hierarchical organization of annotated data files within a publication supplemented by the ability to use identifiers as metadata values to specify relationships between data files and with a network of people, publications, projects, funding sources, or other relevant entities. In essence this allows each group using SEAD to create a custom data model that may have a complex , multi-relationship graph of related data files with annotations spanning standard vocabularies such as Dublin Core, domain vocabularies such as CUAHSI's Observational Data model (ODM), and custom terms.

SEAD as an Exemplar

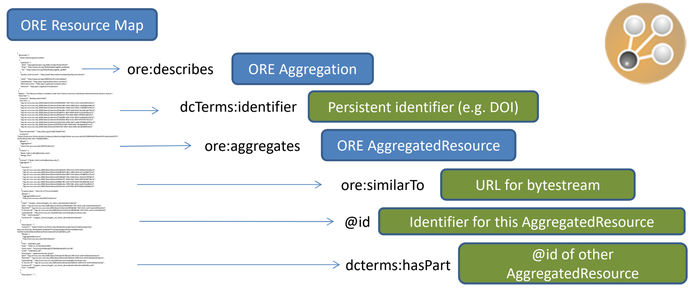
Data Publication across Projects, Data Models, and Repositories:

Combined with SEAD's ability transfer datasets to third-party repositories, which have their own internal data models, the support for user customization requires SEAD's publication services to be very general, with minimal required structure, and a simple, standard means of serializing custom information. SEAD has also had to support migration of data between its version 1 and version 2 services, which differ in the specifics of their data models. While both models are hierarchical, version 1 supported hierarchical Collections containing Datasets that refer to one file and allowed Collections and Datasets to have multiple parents, while version 2 supports Datasets that contain hierarchical Folders and Files, where each item has a single parent, along with a hierarchical Collection mechanism that can be used to group Collections and Files. SEAD's need to address interoperability internally has made it a useful exemplar for the broader interoperability desired across NDS.

SEAD's Publication Approach:

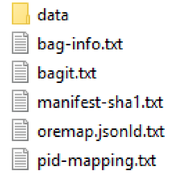
RESTful API with JSON-LD Content: SEAD is architected to support multiple independent sources using publishing services to interact with independent repositories to publish richly structured and annotated datasets. A RESTful API has been created to support create-read-update-delete (CRUD) operations to manage information about repositories, people, and publication requests. These RESTful services exchange JSON-LD formatted content. JSON-LD retains the readability that has made key/value-pair-oriented JSON so popular while adding the concept of an '@context' object in which simple keys such as "Title" can be associated with formal vocabulary terms (e.g. http://purl.org/dc/terms/title) to provide clear semantics and support mapping to RDF.

OAI-ORE: In SEAD, data publications are described using the Open Archives Initiative Object Reuse and Exchange standard (https://www.openarchives.org/ore/), serialized as JSON-LD (http://www.openarchives.org/ore/0.9/jsonld).



SEAD's Data Publications are an *Aggregation* described by a *Resource Map* that contains a flat list of *AggregatedResources*. Repositories are required to assign a persistent identifier to the *Aggregation*. Each *AggregatedResource* is given a unique tag-style persistent identifier and, for those representing files/bytestreams, is associated with (*similarTo*) a URL from which the bytes for the resource can be retrieved. Additional structure is encoded as relationships based on resource identifiers, with the Dublin Core Terms 'hasPart' relationship being used to document SEAD's default folder/file hierarchy.

Reference repository and Library of Congress BagIT serialization: SEAD does not define how repositories make data available online or in serialized form. However, to support large publications (~1TB, 500K+ files) and serve as a reference, SEAD has created a reference repository that provides a landing page, and URLs to download the Resource Map, individual files, and a serialized form of the complete publication. The data publication is serialized as a single Zip file, formatted according to the BagIT specification (https://tools.ietf.org/html/draft-kunze-bagit-11), which defines several standard metadata files providing a basic description of and fixity information for the publication, along with a requirement that data files be in a 'data' subdirectory. To incorporate the additional metadata available from SEAD, we follow the convention developed by DataOne (http://bit.ly/2oYQwtv) to include and name the ResourceMap file and provide mappings between the identifiers used in the map file and the paths used for those resources within the Bag.



Towards Universal Accessibility

The model and metadata agnostic mechanisms adopted on SEAD have been very effective in enabling SEAD to handle customizations for groups in disciplines ranging from hydrology to the humanities, to publish data to repositories based on DSpace, Fedora, and flat files, and to leverage its standard publication mechanism to transfer content between our version 1 and 2 software with their differing data models. While SEAD has not fully addressed the goals of the NDS UADP pilot, our experiences to date, and analysis of how the directions being discussed by UADP participants could be adopted by SEAD, has relevance for other projects interested in helping to define a UADP approach and/or adopting it once it exists:

- **A standard mechanism, given a data publication identifier (in any scheme) to resolve the available metadata:** Nominally this task involves standardizing how to discover a URL from which metadata can be retrieved (or otherwise obtaining a metadata document) and how to interpret the document. For discovery, SEAD has did not find a relevant standard but, given that it does provide both a human-readable landing page and a URL for retrieval of the Resource Map, it should be straight forward to adopt the convention being discussed in UADP to leverage HTTP Accept headers as shown:

  For interpretation, SEAD has found the use of OAI-ORE and JSON-LD as a basic framework to be valuable and flexible. Combined with a "dumb-down" approach (http://wiki.dublincore.org/index.php/Glossary/Dumb-Down_Principle), in which the receiver interprets what it can and retains and displays what it doesn't using simple heuristics, has enabled SEAD to seamless handle publications with significant differences in structure and annotation. While UADP is still discussing options in this area, it is encouraging that several projects including DataOne, the Data Federation Consortium, Hydroshare, and Globus are all using or exploring the use of OAI-ORE.
- **A standard means of harvesting available data publication identifiers from a given repository:** SEAD's services can provide a complete list of data publications created through SEAD and, with minimal work could filter those by date, project, creator, or other basic bibliographic information. Although the UADP group has not yet had much discussion in the area beyond identifying relevant existing work, from SEAD's perspective, having a separate mechanism to retrieve the full metadata should help minimize the detail required in such a listing and hopefully simplify achieving consensus and implementing a solution.
- **Documentation of the current practices:** SEAD has analyzed metadata entries across all of its users' collections and, while there is variation and customization, established standards such as Dublin Core and W3C provenance account for a large fraction of use, suggesting that best-practice guidance may be possible across NDS as well and that de facto minimal schema may exist.
- **A standard means to discover and retrieve individual files within a data publication:** SEAD's convention of using dcterms:hasPart to indicate a default hierarchy and ore:similarTo to indicate which entries correspond to files has proved sufficient to allow data to be ingested by multiple repositories and has simplified migration of content between SEAD's two versions. Some variant of this convention could be leveraged to provide universal accessibility across UADP participants, but the variation seen in SEAD in the semantics of file and path names suggests that consensus beyond this, e.g. to identify 'raw' versus 'final' data or otherwise interpret the provided hierarchy, may be difficult.

Conclusions

Through the SEAD project and discussions to date within NDS on interoperability and universal accessibility, I have become convinced that researchers' ability to discover and use data across different cyberinfrastructure components can be significantly increased without the level of harmonization of data models and metadata (terms as well as allowed values) that would be required for full automation and complete consistency across systems. As the UADP Pilot group has moved from discussion, to documenting possible solutions, related standardization efforts, and current practices in operational cyberinfrastructure, and to a concrete proposal in support of it's Task 1, it has been encouraging to see that there is significant commonality across systems already and that the type of model-agnostic approaches outlined above appear sufficient to manage the differences.
The UADP Pilot group will be meeting at the NDS 7 Workshop and new participants are welcome to join the effort. There project's wiki site includes a broad range of information and additional contributions, including information on relevant standards, current practices, specific use cases, current practices, and/or proposals related to any of the group's four tasks, from any participants are very welcome .

For More Information

SEAD: http://sead-data.net/
NDS UADP Project: https://nationaldataservice.atlassian.net/wiki/display/NDSC/Universally+Accessible+Data+Publications+Pilot

Acknowledgements

The author acknowledges support through the SEAD project, funded by the National Science Foundation under Cooperative Agreement #OCI0940824.
The author also acknowledges:

- The SEAD team from the University of Michigan, University of Illinois, and Indiana University (listed at http://sead-data.net/about/sead-team/) for the collective effort to develop SEAD's services and for discussions related to data publishing and interoperability, and
- The National Data Service Consortium members who have participated in interoperability discussions at prior workshops and the individuals and projects that are participating in the Universally Accessible Data Publications Pilot project for their work in formulating the UADP project scope and the ongoing effort to achieve its goals.