Project Summary:

Name of project:

Equipping OPeNDAP with data citation functionality

Project lead and contact details:

Niklas Griessbaum <griessbaum@ucsb.edu>

James Frew <frew@ucsb.edu>

Proposed start and end date:

2018-02-01 to 2019-01-31

Budget Requested: $4,300 (estimate)

Budget Summary:

- $2,000: 2 persons to attend Summer 2018 meeting (no air travel required)
- $2,000: 1 person to attend Winter 2019 meeting
- $0,300: cloud instance (6 months, 50 USD/month; 1 CPU, 4 GB RAM, 500 GB storage)

This project relates to the following key priority areas (please indicate all that apply):

☒ Earth Science Cyber-infrastructure

□ Semantic Technologies

□ Socioeconomic value of data

□ Other (explain)

Project Outline:

Project description:

Data citations are persistent and technology-agnostic interfaces, serving multiple purposes: They give credit, provide provenance, and facilitate the re-use of data. Properly implemented data citations should therefore promote data sharing, increase transparency, and allow researchers to extend each other's work.

In contrary to printed materials, data is often published as a dynamic service as opposed to a static dataset; it may evolve over time and have varying authorship for different parts of the dataset. Furthermore, data structures are often richer than the document hierarchies supported by conventional citations. This is particularly an issue for geospatial data. Because typically subsets, rather than whole datasets are used, citation of data can become complex to handle.

Since there is no standard way to cite data, most current data citations are really just citations to papers or websites that describe a dataset. They are created by hand and often refer to entire datasets or copies of their subsets. This approach is prone to errors, does not scale in an acceptable way, and becomes problematic when data is updated or appended. The lack of an automated and convenient way to generate accurate citations of data stands in the way of a widespread use of data citations.

We have begun to address this problem by extending the API of the Open-Source Project for a Network Data Access Protocol (OPeNDAP) to generate citations that precisely match the requested data. We have implemented a prototype that allows software and users to query OPeNDAP resources for citations alongside the actually requested data.

We propose to continue the development of this prototype and to expose it to the public for testing and feedback.

Project objectives:

The objectives of this project are to add features to the prototype of the OPeNDAP data citation generator and to port the prototype to a system that is publicly available for testing. The added features will address the user interface and the ability to use citations to retrieve cited data. The porting efforts will address robustness and security. Once available for testing, we will collect feedback from the community on the system. By collecting feedback, we hope to identify use cases and required features. By the end of the project, we intend to have resolved following challenges:

Metadata: Citations will be be created from metadata that currently is exposed through OPeNDAP's data attribute structure (DAS). Therefore minimal requirements on the metadata and robust rules of how to create citations from the metadata will be developed. Mechanisms for dealing with missing metadata (e.g. an external metadata repository) will be explored.

Fixity: Since data may change over time, mechanisms to identify and distinguish between states of the data will be established to ensure that citations point to the actually cited state of the data. We will investigate ETags (e.g. last modified time of a dataset), fingerprints (e.g. cryptographic checksums of the cited subset), or assigned fixity (e.g. versions).

Description of key project steps and timeline:

Part 1:

We will extend the prototype with a user interface and the ability to use citations to retrieve cited data. We will prototype alternative fixity representations. We will demonstrate the system at the Summer 2018 ESIP meeting.

Part 2:

We will port the prototype to a publicly accessible system hosted on an ESIP cloud server for the remainder of the project. We will address robustness and security. We will collect feedback from the community, for presentation at the Winter 2019 ESIP meeting.

Benefits and advancements of key Earth sciences priority areas:

We believe that the usage of data citations is necessary to ensure that

- due credit is given to data publishers,
- data provenance is established, and
- data access is facilitated.

These points are the foundation required to help promote transparent sharing of data, which will increase reproducibility and interoperability, ultimately leading to reduced redundancy and increased cooperation. However, we further believe that only through the availability of automated (and therefore convenient) data citation generations will widespread usage of data citations be established. Data citations provided by OPeNDAP may be passed on to the APIs of higher service layers and finally to the application user interface, where they can be ingested by a broad audience of users.

Apart from these direct impacts, the project will serve as a reference example for the integration of automated citation generation into data provision services and may facilitate and encourage the adaptation of automated citation generation in other data services.

Project Partners:

Involvement of ESIP collaboration areas:

The proposed project directly addresses the “data identifiers” activity of the Data Stewardship Committee.

Additional Information:

What groups/audiences will be engaged in the project?

In the first phase, our main audience will be data providers. The objective is to ensure that data providers are sensitized to the necessity of automate citation generation and to provide a solution to them to address this issue.

In the second phase, the audience will be any earth science data user. The objective here is to inform about the benefits of using data citations and the provision of convenient tools to generate data citations.

How will you judge that project has had impact?

Ultimately, the claim of this project is to help increase the use of data citations. More specifically, we hope to see our approach to automatic data citation adopted by OPeNDAP data providers, and/or by other data services.

How will you share the knowledge generated by the project?

The extensions to the “hyrax” OPeNDAP reference implementation will be published as a repository on GitHub. The findings and experience will further be disseminated at ESIP meetings and documented as a journal article.