ADVANCING OPEN SOURCE, LINKED DATA SOLUTIONS FOR CROSS-REPOSITORY DISCOVERY AND COLLABORATION: Putting the WHOAS Community Repository in a Geoscience Context

The GeoLink Project

There are many different repositories that house geoscience metadata, science data, and publications related to the same people, projects, and activities. These repositories are run by many different organizations, with different needs, budgets, priorities, and technologies.

Finding things across these platforms often requires an unreasonable amount of expert knowledge, not just in the field of science, but also in the field of data policy and information retrieval. Answering larger questions often constitutes an entire research project of its own.

**What papers have been published using data collected from a particular area?**

If our repositories could talk to each other, we would have a much better picture of our data in the context of geoscience. But our systems don’t speak the same machine language, much less the language of geoscience. To address this, GeoLink partners are using Linked Data, a rapidly adopted, open source technology that sits on top of our content systems. Repositories can create RDF using Linked Data, a rapidly adopted, open source technology that could be used to expose RDF for many uses—but the goal of our GeoLink work was to end up with RDF that can be integrated with our geoscience partners.

DSpace Development

The library has been collaborating with @mire, Inc., a registered DuraSpace service provider, to make our DSpace repository Linked Data optimized.

**DSpace 5 is a big step towards Linked Data, but we needed to go further.**

First, we needed to change the way DSpace thinks. For example, Person is not an “exact string”, a Person is a concept with a URI and multiple possible string variations. The Person concept is the most obvious example in DSpace, but the mechanism works for any important aspect of your repository, like Location, Organization, and Project.

Another important aspect of development is the ability to map external RDF resources from SPARQL end points into local authority concepts.

Translating Our Linked Data

One of our main goals in developing custom DSpace code, was to end up with a useful, broadly applicable module that could be used to expose RDF for many uses—but the goal of our GeoLink work was to end up with RDF that can be integrated with our geoscience partners.

To achieve the integration piece, we worked with Adam Shepherd, an experienced WHOI developer, to create a script that queries our SPARQL endpoint for data, and rewrites the RDF to conform to GeoLink patterns.

The Reward

- **Sharable, Customizable DSpace Code**
  An open source solution for libraries that want to use and produce Linked Open Data as part of their DSpace repository

- **Leveraging SPARQL Endpoints to Improve Metadata**
  User-configurable, authority-controlled fields that link out to external data sources and preserve data locally

- **A Linked Open Data Repository**
  WHOAS records represented as GeoLink objects that can be searched and harvested across other geoscience collections

Acknowledgments

This work was funded by the National Science Foundation

EAGER: Collaborative Research: Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery
EarthCube Building Blocks: Collaborative Proposal: GeoLink - Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences