UNIVERSITY OF CALIFORNIA

Santa Barbara

Spatial Discovery and the Research Library: Linking Research Datasets and Documents

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Geography

by

Sara Lafia

Committee:

Professor Werner Kuhn, Co-Chair

Professor Krzysztof Janowicz, Co-Chair

Dr. Katja Seltmann

March 2017

The thesis of Sara Lafia is approved.

_____

Katja Seltmann

_____

Krzysztof Janowicz, Committee Co-Chair

_____

Werner Kuhn, Committee Co-Chair

January 2017

ABSTRACT

Spatial Discovery and the Research Library: Linking Research Datasets and Documents

by

Sara Lafia

Academic libraries have always supported research across disciplines by integrating access to diverse contents and resources. They now have the opportunity to reinvent their role in facilitating interdisciplinary work by offering researchers new ways of sharing, curating, discovering, and linking research data. Spatial data and metadata support this process because location often integrates disciplinary perspectives, enabling researchers to make their own research data more discoverable, to discover data of other researchers, and to integrate data from multiple sources.

The Center for Spatial Studies at the University of California, Santa Barbara (UCSB) and the UCSB Library are undertaking joint research to better enable the discovery of research data and publications. The research addresses the question of how to spatially enable data discovery in a setting that allows for mapping and analysis in a GIS while connecting the data to publications about them. It suggests a framework for an integrated data discovery mechanism and shows how publications may be linked to associated data sets exposed either directly or through metadata on Esri's Open Data platform. The results demonstrate a simple form of linking data to publications through spatially referenced

metadata and persistent identifiers. This linking adds value to research products and increases their discoverability across disciplinary boundaries.

Current data publishing practices in academia result in datasets that are not easily discovered, hard to integrate across domains, and typically not linked to publications about them. For example, discovering that two datasets, such as archaeological observations and specimen data collections, share a spatial extent in Mesoamerica, is not currently supported, nor is it easy to get from those data sets to relevant publications or other documents. In our previous work, we had developed a basic linked metadata model relating spatially referenced datasets to documents. The research reported here applies the model to a collection of spatially referenced researcher datasets, capturing metadata and encoding them as linked open data. We use existing RDF vocabularies to triplify the metadata, to make them spatially explicit, and to link them thematically. Our latest research has produced a simple and extensible method for exposing metadata of research objects as a library service and for spatially integrating collections across repositories.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 1. Spatial Discovery and the Research Library

Sara Lafia, Jon Jablonski, Werner Kuhn, Savannah Cooley, F. Antonio Medrano

I. Introduction

Location plays a key role in the organization and integration of knowledge. In an interdisciplinary setting, location can reveal patterns and trends in diverse and seemingly disparate information. For example, a "geographic prism" on social mobility data in the United States reveals vast regional differences that can then produce hypotheses about causes, based on local differences in factors like family structure or schools ("Mobility, measured", The Economist 2014). Data discovery tools that exploit location can offer users a spatial view of phenomena, and in doing so, bridge disciplines in research and policy making. The design of such tools, with an emphasis on connecting the discovered data to publications about them, is the focus of this paper.

A. Problem Statement

Enabling the spatial discovery of research publications and datasets, herein referenced as research objects, is the next step in the evolving role of the modern research library. The notion of the extensible and reusable research object originates from the domain of e-Science (Bechhofer et al. 2010). Over time, the set of research objects, beginning with documents, has expanded beyond texts to include artifacts, models, games, and works of art (Buckland 1997). Today, research libraries are increasingly called upon to build links between research objects, such as journal articles or electronic theses, and auxiliary data, of which both may have embedded locational references and may reside in external data repositories.

Emerging research object repositories, which hold data and publications, are still unstable as architectures and face challenges in handling spatially referenced content (Hey et al. 2009). There is a growing need for a stable yet flexible discovery mechanism that can thrive in an evolving spatial and non-spatial information landscape (Cooley et al. 2015). At the same time, e-Science is producing sophisticated models of research objects (Bechhofer et al. 2010) that are exceedingly complex for the needs of data and publication discovery at libraries. Much work remains to be done in the development of simple search and discovery tools that span multiple collections (van Hoolen et al. 2014).

In this article, we address the primary challenge of stability by implementing a simple linked data model that exploits basic relationships between research data and research publications in a way that does not break when repositories change. In doing so, we also address a second challenge, that of supporting discovery, resulting in enhanced integrative capacity for spatially referenced research objects. Combining these two challenges in the proposed pragmatic form is expected to result in progress on a third, broader goal: that of supporting interdisciplinarity in scientific workflows through data reusability, within and across domains. These three challenges translate into the following set of guiding research questions:

1. How can libraries generate stable links for research objects across repositories?
2. How can libraries support the discovery of research objects based on location?
3. How can libraries promote cross-disciplinary data sharing and reuse?

The research, undertaken by the Center for Spatial Studies at the University of California, Santa Barbara (UCSB), in partnership with the UCSB Library and Esri Inc., seeks to make spatial references and relationships explicit in research objects, thereby

integrating diverse contents and contextualizing published research data by connecting them to publications.

B. Motivation

Research institutions generate massive quantities of data from diverse disciplines in a wide variety of formats (Mayernik et al. 2015). Recent efforts to increase transparency and reproducibility encourage, and often mandate, that researchers make publications and data publically accessible through open-access licenses (University of California Regents 2014). A proliferation of associated data is contributing to a growing imbalance between an institution's ability to collect data and its ability to curate resources (Cragin et al. 2010), resulting in a trade-off between quality assurance and ingestion capacity, a trend that the UCSB Library can attest has accelerated in the intervening years.

Interdisciplinary research presents additional unique challenges, including disciplinary differences in frames of reference, operational agendas, research methods, and vocabularies (Brewer 2015; MacMillan 2014). Many data discovery portals support only domain-specific vocabularies, data structures, and metadata formats, severely limiting the applicability and reuse of data across domains (Golding 2009). More limitations arise when subsets of domain research are published in expensive subscription journals. This diminishes research impact and potential for data reuse across domains, which could be enhanced if made available through open-access policies (Harnad et al. 2008).

Library repositories have developed detailed workflows for generating metadata that use rigorous metadata content standards and controlled vocabularies, but result in extremely limited ingest capacity. In contrast, repositories for self-deposit of spatial content, such as ArcGIS Online, have minimal metadata constraints and see 8,000–12,000 new and mostly

undescribed objects added per day (Szukalski 2015). Metadata that describe the lifecycle of a dataset are often very granular and must account for both library and spatial needs. Metadata are valuable for long-term preservation, yet they are not central to resource discoverability (Hardy and Durante 2014), which is the primary focus of this research. Enforcing particular metadata requirements may do more to hinder data availability, especially across diverse domains that have their own metadata standards, than to aid in their discovery.

Library-run and self-deposit systems of research data management can be complementary, as they approach control and sharing in two distinct ways. However, they are not currently connected. This research proposes to align the traditional library ingest process with the self-deposit approach of cloud-based GIS, such as ArcGIS Online, through the generation of links between two sets of research objects: researcher publications and researcher data. This approach combines the best aspects of both worlds: spatial discovery of data from the GIS world and document curation from the library world, connected through a lightweight and stable linked data solution.

Creating links between publications held in a tightly controlled library repository and data stored across external databases increases the discoverability of these research objects. This work connects research objects held in separate repositories without the need to formally align their metadata schemas. In our proposed framework, a research object in a self-deposit environment like ArcGIS Online[1], which has minimal metadata constraints, is semantically linked to a related research object in an institutional environment with tightly controlled metadata.

[1] https://www.arcgis.com/home/

This article presents a proof-of-concept model for linking spatial data to the research publications that utilize them. Using OpenRefine with its Resource Description Framework (RDF) extension for data processing and cleaning[2], we link sample publications to data hosted on Esri's Open Data platform by Dublin Core metadata relationships. The linked data are stored as triples, which allows for queries on the associated RDF about publication data. Such formalized relationships are key to developing a rich publication and data repository that allows for discovery of research resources and advances cross-disciplinary sharing of knowledge, as illustrated in Figure 1.

Figure 1. Project vision for data discovery and publication integration across domains.

II. Background and related work

This work builds on a long tradition of spatially enabled digital libraries and uses the latest semantic and geospatial technologies to demonstrate the potential for spatial discovery and the interlinking of research resources. As university researchers are increasingly

2 http://openrefine.org/download.html

expected to share the data associated with their publications under open data mandates, university libraries find themselves being called upon to curate increasing volumes and additional types of researcher-generated data. In this context, enhancing users' ability to share, discover, and make sense of content is of great importance.

A. Library repositories

In the mid-1990s, the Alexandria Digital Library (ADL) at UCSB was the first distributed digital library (Freeston 2004) to offer collections of georeferenced materials, hosted online, searchable by spatial and temporal criteria (Goodchild 2004). ADL eventually lapsed, relegating UCSB researcher data, such as the popular Maya Forest GIS collection (Ford 1995), to offline discovery and curation. Reinstating such legacy collections through an open-access digital presence increases their utility in an interdisciplinary research context. Further, linking these datasets to publications in a manner that can be exploited by Semantic Web tools improves their discoverability.

Many university libraries have implemented hybrid ad-hoc solutions for spatial data collection, discovery, access, storage, and archiving in the context of the changing landscape of user needs and technologies (Scaramozzino et al. 2014). Libraries have generally promoted interdisciplinary collaboration by supporting geospatial research platforms and tools for analysis and post-data discovery. However, they do not yet combine spatial and semantic approaches to expose connections between existing data silos that span diverse disciplines. In practice, most library-curated research objects are locally stored, have limited access points, and are undiscoverable from related content (Padilla 2016).

UCSB Map & Imagery Laboratory (MIL) is in the process of developing a spatial metadata workflow using ArcCatalog for the purposes of preparing spatial datasets for

ingest into the new Alexandria Digital Research Library (ADRL)[3]. The ISO 19115 standard[4], the Open Geoportal Metadata Creation Guide[5] and the Stanford University metadata creation workflow[6] inform this metadata model. The work described in this paper couples these ongoing efforts with the production of linked data.

B. Emerging spatial data technologies

Achieving the dual purposes of enhancing spatial discovery and linking research objects requires a novel solution. Some contemporary data management solutions address the need to enable the spatial discovery of resources, but do not enhance discovery of resources through semantic links. GeoBlacklight[7] for instance, is an open source, multi-institutional software project that many libraries are currently adopting (Addison et al. 2015; Durante and Hardy 2015). It offers users text-based, spatial, and faceted semantic search to enable discovery of GIS-consumable resources across organizations (Hardy and Durante 2014). GeoBlacklight also allows users to connect to data as a service, which enables analysis from a desktop GIS, comparable with Esri Open Data. While a GeoBlacklight instance for the UCSB Library would support spatial discovery by relating spatially referenced content based on location, it would not connect the data to publications held in the library's own repositories, nor to other external repositories. Another possible data management solution includes the California Digital Library's Dash system, which has been adopted by several University of California campuses and features a self-deposit feature, facilitates data search,

3 http://alexandria.ucsb.edu/

4 http://www.iso.org/iso/catalogue_detail.htm?csnumber=53798

5 http://opengeoportal.org/working-groups/metadata/metadata-creation-guide/

6 https://lib.stanford.edu/metadata/documentation

7 https://github.com/geoblacklight/geoblacklight

data sharing, and preservation services (Tsang 2015). However, DASH does not offer inherent spatial functionality, although efforts to achieve this are underway at UC Irvine[8].

Considering these existing alternatives, utilizing Esri's ArcGIS Online platform as a foundation for combined spatial and semantic search makes sense for several reasons. Since GIS software has become ubiquitous for performing spatial analysis across a variety of academic disciplines, universities often administer an ArcGIS Online enterprise account through their libraries. ArcGIS Online is a cloud-based GIS that acts as a self-deposit data system with basic geoprocessing functionality. Additionally, ArcGIS Online now includes Esri Open Data , which is a spatial data repository with native access controls and search features. Enabling Open Data on ArcGIS Online allows organizations to make content available to the public or restricted to users authorized by the institution.

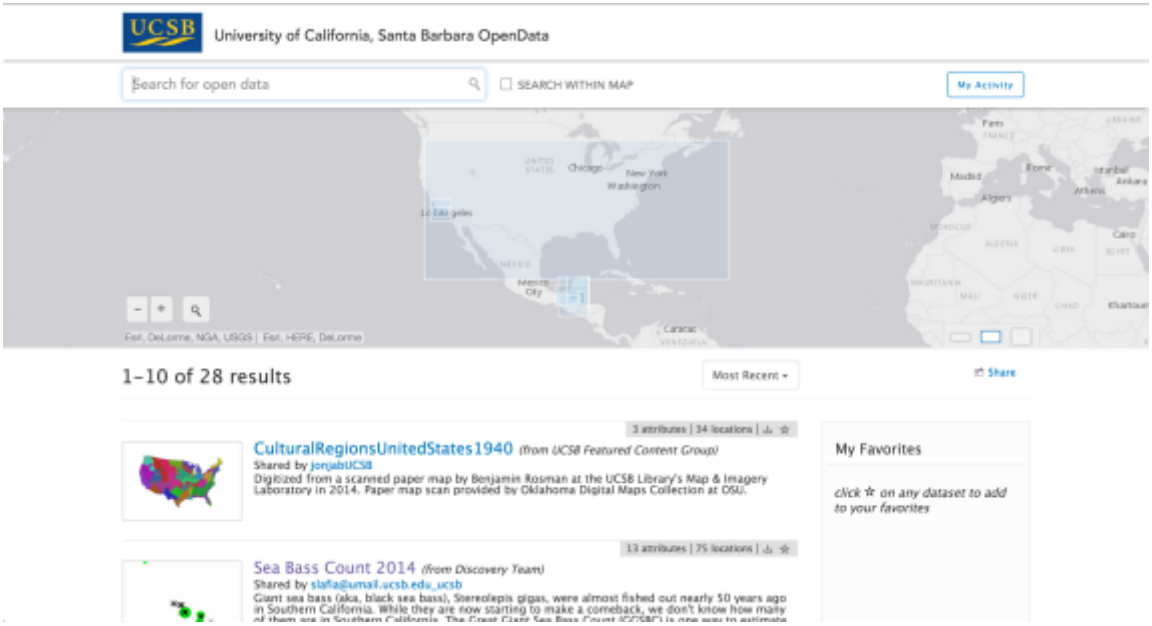
There are many advantages to using Esri Open Data as a spatial data discovery solution, not the least of which is publishing spatial data in a way that allows open access and download. Users are not required to have ArcGIS Online credentials to access data hosted through Esri Open Data[9], which increases both accessibility to data and reproducibility of results derived from that data. ArcGIS Online also supports various metadata standards, increasing the potential to share data across domains. Its interface allows for visualization and filtering of the data for basic geoprocessing and analysis. This adds immediate value to the discovery process, as users can begin making sense of datasets even before downloading them. Many organizations are adopting Esri's Open Data platform because ArcGIS Online

8 https://dash.lib.uci.edu/xtf/search

9 http://opendata.arcgis.com/

offers web-based analysis and search. UCSB's instance of Esri Open Data[10] is shown in Figure 2.

Figure 2. UCSB's Open Data instance leverages ArcGIS Online.



While Esri's Open Data platform is an excellent tool for publishing, discovering, and accessing spatial datasets, it is not a stable repository solution in either the traditional sense of institutional preprint repositories or in the emerging sense of Trusted Digital Repositories (Tsang 2015). However, when linking data with a controlled resource, such as UCSB's Alexandria Digital Research Library[11] (ADRL) repository, which hosts theses and dissertations, or University of California's eScholarship[12], which offers open-access to researcher publications, the power of the Semantic Web can be brought to bear on the systems. This design choice provides flexibility that many current repositories cannot offer.

10 http://discovery.ucsb.opendata.arcgis.com/
11 http://www.alexandria.ucsb.edu/
12 http://escholarship.org/

C. State of the art

Many institutions, including libraries, archives and museums, are adopting linked data approaches to improve the discoverability of the growing number of resources that they curate (van Hooland et al. 2014). Institutions, such as the Linked Data for Libraries university consortium, the Library of Congress and the Tate Modern Gallery, leverage linked open data technologies to enhance access to their collections. These management models offer users access to data and metadata through Application Programming Interfaces (APIs) and extend query capabilities through accessible endpoints.

The Linked Data for Libraries (LD4L)[13] initiative is a multi-institutional effort, including Stanford, Rice and Harvard universities, aimed toward applying the Library of Congress Bibliographic Framework Initiative to describe library resources. Transforming traditional MARC (MAchine-Readable Cataloging) metadata descriptions, which are flat, text-based, and fielded (Avram 2003) into linked BIBFRAME descriptions for cartographic and geospatial materials leverages Library of Congress controlled vocabularies alongside DBPedia and GeoNames, to model places, creators, themes, and events (Durante et al. 2016). Library of Congress is a notable early organizational contributor to the production of API-accessible linked open data for authority files[14]. Many institutions use these services for authoritative reconciliation (Heath and Bizer 2011). These services allow institutions, such as the Tate Modern Gallery, to contribute collection metadata to repositories, like GitHub[15], that they neither own nor manage, increasing discoverability, content exposure and creative reuse (Padilla 2016).

13 https://www.ld4l.org/

14 http://id.loc.gov/

15 https://github.com/tategallery/collection

The development of Semantic Web technologies enables linked data driven portals. Linked data portals provide new opportunities to organize metadata and retrieve information resources such as text documents, datasets, and multimedia content (Baierer et al. 2014; Hu et al. 2015-1; Hu et al. 2015-2). Linked data resource discovery systems can index domain-specific information with terms from ontologies. Ontologies are formal explicit specifications of a shared conceptualization using a vocabulary of classes and relations, expressed in RDF, which is a data model that stores metadata attributes as nodes and links to constitute an interconnected graph.

Whereas other methods for publishing data rely on multiple data models, the RDF data model provides an integrated and simple access mechanism that also supports hyperlink-based data discovery using uniform resource identifiers (URIs) as global identifiers for entities (Heath and Bizer 2011). For instance, Athanasis et al. (2009) described data with domain-specific spatial ontologies in a linked data discovery tool, and Keßler et al. (2012) developed a linked data portal for the GIScience community to explore and visualize geographic distributions of publications by conference location and editor or author affiliations. Scheider et al. (2014) have leveraged linked spatiotemporal data to enhance access to diverse formats of library materials, from paper maps to scientific datasets. Taken together, the interlinking of research objects and their metadata creates a semantically linked graph.

Adopting semantic technologies addresses issues of interoperability that arise from online portals featuring spatial data in various standards and formats. In particular, relationships between research publications and associated data can be captured through

RDF subject-predicate-object triples, which bridge gaps between data and metadata, as well as differing metadata content standards.

III. Methods

Publicly available research objects, namely researcher datasets and researcher publications, drive the data discovery mechanism developed in this research. The design and evaluation of a linked data model is informed by user personas, which structure the relationship between published research and associated data. The extensible triple model developed in this work allows for future expansion of the vocabulary.

A. User personas

The current designs of most access systems do not support the spatial integration of research object collections across various domains. Adopting the personas of domain scientists and considering the types of data that each might search for or contribute, along with their motivations for doing so, informed the design specifications of our system. The UCSB Esri Open Data instance contains collections of test data that span research domains, data formats, and user needs. The current three exemplary data collections represent a small but diverse range of disciplines, from archaeology to political science, and diverse formats, including shapefiles, imagery, text documents, external repositories, and map services.

The first collection corresponds to Anabel Ford's Maya Forest GIS and was obtained from a CD archive (Ford 1995). The data include shapefiles and imagery complete with full ISO compliant metadata created by UCSB Library staff. The second collection comes from a meta-analysis conducted by Benjamin Halpern, a UCSB ecologist. His collection of sampling sites has a global extent and is hosted in an external repository, a practice typical

of UCSB researchers in the life sciences for disseminating research (Halpern et al. 2009). While these spatial data are open-access and publically shared, they are not currently discoverable through a search of UCSB Library holdings. The third data collection comes from Thomas Patterson, a political scientist at Stanford University, and represents world boundaries of disputed areas. The data are part of the broader Natural Earth collection, currently discoverable through the UCSB Library, but not yet formally associated with Patterson's research publications (Patterson 2009).

Table 1. Personas, domains, and datasets of researchers currently discoverable through UCSB Open Data.

Persona	Domain	Dataset	Dataset location	Publication	Publication location
Anabel Ford	archaeology	Archaeological Sites[16]	UCSB Open Data	Assessing Situation El Pilar[17]	UCSB eScholarship
Benjamin Halpern	ecology	Science of Marine Reserves: Meta-analysis[18]	Knowledge Network for Biocomplexity	Biological effects within no-take marine reserves: a global synthesis[19]	UCSB UC-eLinks via WorldCat
Tom Patterson	political science	World Boundaries of Disputed Areas[20]	EarthWorks	Natural Earth[21]	SearchWorks

16 http://opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0

17 http://escholarship.org/uc/item/4qr2x8p3

18 https://knb.ecoinformatics.org/#view/doi:10.6085/AA/pisco_smr_synthesis.1.3

19 http://ucelinks.cdlib.org

20 https://earthworks.stanford.edu/catalog/stanford-tq310nc7616

21 https://searchworks.stanford.edu/view/11047527

The personas, summarized in Table 1, cover various data sharing scenarios. Anabel Ford has locally hosted resources that she intends to share with a global public audience through open access. Benjamin Halpern is from a domain that favors data distribution through a repository external to UCSB. Thomas Patterson is from another institution and has spatially relevant contents that might interest Ford, Halpern, or other scientists at UCSB or anywhere else. The datasets share a spatial overlap that would not otherwise be obvious. For instance, Patterson's contested borders dataset is a feature collection with a global extent, yet intersects with Ford and Halperns' regions of research. The potential to expose the spatial complementarity of resources would go unrecognized without the assignment of spatial footprints to these objects. Users can then benefit from discovering useful and seemingly unrelated datasets or publications from unfamiliar domains by exploring the spatial relations of the research objects.

Taken together, these exemplary researcher personas and associated datasets provide a foundation for several competency questions that capture the kinds of queries that users may want to construct:

- Find datasets referenced by a particular publication.
- Find publications that have a particular dataset associated with them.
- Find research objects that overlap with a particular spatial extent.

Performing such queries is frequently relevant to a resource discovery process, but relationships between research data, the publications that reference them, and the locational extents that they cover are not currently exposed in the metadata. The onus of relating publications with datasets, as well as relating both the publications and datasets with

location, is currently placed on the end-user. The linked data relationship between research publications and data, taken along with the spatial extent of the dataset represented in Open Data, address the types of thematic and spatial queries that users would currently like to ask of a library catalog but cannot.

B. Experimental design

The purpose of using linked data in our approach is to formalize relationships between data hosted through Esri Open Data or any other spatial repository, and publications hosted anywhere. The linked data publishing pattern followed in this research generates linked data from static structured data in the manner of Heath and Bizer (2011). This is achieved by taking static input data in the form of spatial and non-spatial contents, publishing them as services and generating a triplestore to reference the URIs of data services and associated publications. This is achieved with the aid of the tool OpenRefine and its RDF extension. The stepwise procedure undertaken to achieve this is summarized as follows:

1. **Data hosting**: Spatial and non-spatial research data are published to a local server and shared via ArcGIS Online as image or feature services, which are shared with the UCSB Open Data group by a system administrator and are made publically referenceable through Open Data source URIs.

2. **URIs**: Identifiers for corresponding publications and dataset services referenced by Open Data content are retrieved from open access document repositories or publisher pages.

3. **Vocabularies**: The OpenRefine with RDF extension generates a graph using the identifiers of publications and research object relationships defined by Dublin Core predicates.

4. **Reconciliation**: The graph is referenced against Library of Congress Subject Headings to enrich users' ability to explore and discover thematically linked content.

5. **Implementation**: Publication-data relationships are serialized as triples that can be queried using the SPARQL Protocol and RDF Query Language.

Because the data described in the previous section are hosted as web services, they are easily referenced through their URIs. Researchers at UCSB can currently share their spatial and non-spatial research data through the institutional instance of ArcGIS Online. Any content currently available through this platform can be migrated into Open Data by changing system permissions. A small subset of data are currently hosted for this research, but by hosting data directly on ArcGIS Online and by connecting additional external resources associated with UCSB researchers to UCSB Open Data, we hope to expand content.

The URIs of research objects correspond to either a data layer or a publication. Datasets that share a common base name are parts of collections and are indicated by URI container. Data creators have only partial control over assignment of the URI resource domain name, which as a best practice, should be self-descriptive and human readable (Heath and Bizer 2011).

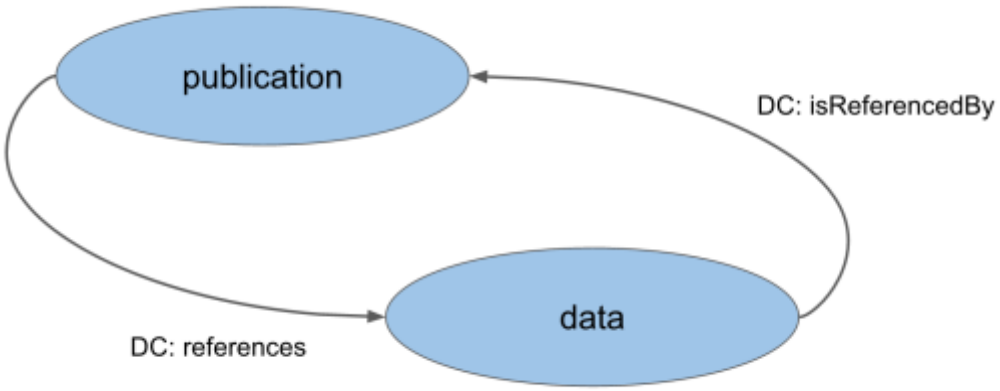
When selecting a technology for RDF creation, it was important to consider the provided data formats, mechanisms of access and desired output. While initial stages of this research tested the Callimachus linked data application builder, a locally hosted triplestore was deemed to be inefficient and limiting. Several other RDF converters and services were considered, but many of these tools perform script-based extraction, transformation and

loading from web pages. Semi-automatic RDF creation, rather than script-based extraction for instance, is a technique better suited to our purposes.

The nature of the data and the questions asked about the data determine the choice of vocabulary. Using predicates from existing vocabularies increases data interoperability and

Figure 3. Generic Dublin Core Metadata Initiative (DCMI) data model.

reuse. Other datasets and applications that use shared vocabularies can also be more readily cross-linked without additional processing, increasing their discoverability (Heath and Bizer 2011). The Dublin Core Metadata Initiative (DCMI) vocabulary is widely used and is well maintained with dereferenceable URIs that point to a retrieval protocol. These factors motivated the decision to use DCMI instead of specialized vocabularies, which are typically less stable. DCMI metadata elements define general attributes such as title and subject. Our data model does not rely on metadata standards, but rather on the two simple associative relationships, *isReferencedBy* and *references*, defined in the Dublin Core ontology[22] shown in Figure 3.



[22] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#

One of the motivations for producing linked data is to forge associations with other data sets, which is a step achieved during the reconciliation process. URIs of the research objects can be interlinked with Library of Congress authority files[23] and even extended to link with other contextually relevant ontologies, such as Wikipedia's knowledge graph DBPedia[24], by referencing the SPARQL endpoints. These links enable exploration of other works associated with authors and datasets.

Once the linked data model has been applied to the publications and dataset URIs, OpenRefine generates an RDF skeleton. The interface allows users to preview the RDF

Figure 4. Reconciled OpenRefine template (above) and RDF skeleton (below).

schema and manually edit nodes in the graph. Once the structure is formalized, it is possible to export the data to a variety of formats, such as RDF/XML or Turtle, depending on the intended use, as shown in Figure 4.

23 http://id.loc.gov/

24 http://mappings.dbpedia.org/server/ontology/classes/

We used OpenRefine with its RDF extension to implement our simple linked data model. OpenRefine generates a static profile triplestore, which is an internally hosted RDFa document that references the URIs assigned to the publication and research data. These static files can then be uploaded to a web server, offering users a web-accessible interface that supports queries.

In OpenRefine, a class is a set of RDF resources that use the same templates. Classes such as publications and data are defined as instances. A new Publications class template

Figure 5. RDF triples for datasets and publications exported in Turtle syntax.

uses an RDFa serialization, embedding RDF as triples in HTML documents and encoding the semantic properties and relationships captured in Figure 5.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .


<http://escholarship.org/uc/item/4qr2x8p3> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:references <http://opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0> ;
        dcterms:URI <http://escholarship.org/uc/item/4qr2x8p3> ;
        rdfs:label "Assessing the Situation El Pilar" ;
        foaf:name "Anabel Ford" .

<http://opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:isReferencedBy <http://escholarship.org/uc/item/4qr2x8p3> ;
        dcterms:URI <http://opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0> ;
        rdfs:label "Archaeological Sites" ;
        foaf:name "Anabel Ford" .

<http://ucsb.worldcat.org/oclc/429112939> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:references <https://knb.ecoinformatics.org/#view/doi:10.6085/AA/pisco_smr_synthesis.1.3> ;
        dcterms:URI <http://ucsb.worldcat.org/oclc/429112939> ;
        rdfs:label "Biological effects within no-take marine reserves: a global synthesis" ;
        foaf:name "Benjamin Halpern" .

<https://knb.ecoinformatics.org/#view/doi:10.6085/AA/pisco_smr_synthesis.1.3> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:isReferencedBy <http://ucsb.worldcat.org/oclc/429112939> ;
        dcterms:URI <https://knb.ecoinformatics.org/#view/doi:10.6085/AA/pisco_smr_synthesis.1.3> ;
        rdfs:label "Science of Marine Reserves: Meta-analysis" ;
        foaf:name "Benjamin Halpern" .

<https://searchworks.stanford.edu/view/11047527> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:references <https://earthworks.stanford.edu/catalog/stanford-tq310nc7616> ;
        dcterms:URI <https://searchworks.stanford.edu/view/11047527> ;
        rdfs:label "Natural Earth" ;
        foaf:name "Tom Patterson" .

<https://earthworks.stanford.edu/catalog/stanford-tq310nc7616> a dcterms:BibliographicResource , dcterms:URI ;
        dcterms:isReferencedBy <https://searchworks.stanford.edu/view/11047527> ;
        dcterms:URI <https://earthworks.stanford.edu/catalog/stanford-tq310nc7616> ;
        rdfs:label "World Boundaries of Disputed Areas" ;
        foaf:name "Tom Patterson" .
```

The triplestore can be queried using SPARQL. A SPARQL endpoint is a web-protocol to which queries against a triplestore can be submitted (Powell 2014). User queries pertaining to datasets referenced within a publication or publications that utilize a particular dataset can be formulated in this way. General queries across all relationships as well as between specific publications or datasets can then be generated. A SPARQL query for all publications that reference datasets can be formulated against the triples, as shown in Figure 6.

Figure 6. A generic SPARQL query against the triples.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT ?data ?publication ?name ?label

WHERE {
    ?publication a dcterms:BibliographicResource
    ?publication dcterms:references ?data .
    FILTER (?name, "Ford")
}

ORDER BY desc(?label)
```

In this query, a user requests the attributes of data associated with publications, which are then optionally filtered by matching author name and sorted by title. By formalizing the relationships between subjects and objects through the use of DCMI prefixes during the data production phase, it is possible to map relationships between research publications and datasets. In this example, matching triples for all publications referencing datasets produced by Dr. Anabel Ford are returned, sorted by title. Additional queries could be constructed using any combination of predefined attributes and predicates.

IV. Results

The research datasets tested in the model included a Maya Forest GIS layer featuring archaeological sites on UCSB Open Data as the object and a published report from the researcher on the 2000 Field Season[25] as the subject (Ford and Wernecke 2000), which are

[25] http://escholarship.org/uc/item/4qr2x8p3

illustrated in Table 2. Queries for datasets associated with a particular publication use the DCMI predicate references to point users to linked datasets hosted through UCSB Open Data. Conversely, users can query for publications associated with datasets through the predicate isReferencedBy, which points back to objects in their respective repositories using URIs.

Table 2. Example of a triple stored in the RDF framework.

Resource	Subject	Predicate	Object
data	Archaeological Sites Maya Forest GIS	isReferencedBy	Assessing the Situation at El Pilar
publication	Assessing the Situation at El Pilar	references	Archaeological Sites Maya Forest GIS

The two parameters defined within the OpenRefine template include a publication URI resource, which is provided by the user, and a data URI resource, which in the case of the sample data comes from Esri Open Data. The relationship between these entities is manually defined. The template references the DCMI vocabulary and makes assignments to each resource based on the user asserted relationship. OpenRefine with RDF extension offers a flexible template that can easily be extended to include additional prefixes and connect the research objects to other collections.

Once linked data are generated from the research objects, publications and datasets are discoverable from their URIs. Users can spatially browse for datasets through the UCSB Open Data instance and discover linked datasets based on the associated attributes formalized in the data model. Importantly, this process enables the spatial discovery of research publications associated with spatial datasets, which are not traditionally

conceptualized as objects with footprints. Retrieving datasets from publications is also possible through the linked data model, as pointers to the hosted data can be exposed during a search on an external repository.

We have deliberately avoided developing a complex model of authorial relationships between data and publications. With a data model for generating simple associative triples in place, scaling up the number of resources referenced in the system from our current small set will be possible.

User-testing to ensure the data model is adequate, and achieving a critical mass of datasets and associated publications will eventually result in a cross-disciplinary discovery resource.

V. Discussion and Conclusions

This article presents a first step in establishing a linked data discovery mechanism that prioritizes stability and supports the discoverability and reusability of research data with spatial references, whether these are in the data themselves or just metadata. It demonstrates how academic libraries can spatially enable the discovery of research objects across disciplines and systems. Formalized relationships between publications and researcher-generated data expose the interplay between researchers and the data that they use or produce. Linking research data, hosted for example through Esri Open Data, to publications, such as those accessible through the UCSB Alexandria Digital Research Library repository, adds value to both sets of research objects. By creating links through the use of linked data predicates taken from Dublin Core, library users are led from publications to data and back, leveraging the spatial search in Esri Open Data on a much broader scale. Making an

increasing amount of content compatible through a linked data model will make more library holdings discoverable through a spatial search interface.

A. Limitations

Current institutional policies support research sharing through open-access licensing, yet incentives and formal channels for sharing only currently exist for publications, not necessarily for associated datasets (University of California Regents 2014). Therefore, in order to lower hurdles to participation, the system described here opts to give researchers full control over what data they want to make available and how.

Another open issue is the long-term maintenance of such a system. The production of linked data is currently a manual process undertaken using OpenRefine. Transitioning to a system that automatically scrapes repositories and generates links may be desirable. The use of semi-automatic RDF creation in this research enabled reconciliation of resources through a graphical user interface, yet this required manual effort. The process could be expedited through the use of server-side tools like Apache Jena[26] to automate the workflow by running periodic scrapes and generating triplestores from URIs.

B. Next steps

Metadata for objects in ADRL recently became available as RDF triples, which are available through a dedicated API. UCSB Library staff harvest metadata in ADRL from the MARC metadata in the library catalog. Aligning this collection with the triplestores for research objects currently generated in OpenRefine can increase the amount of campus resources accessible as linked data, expanding the university's knowledge graph. Linking

[26] https://jena.apache.org/index.html

these systems through common vocabularies could increase awareness of research efforts across domains and increase discoverability of curated research objects

The ADRL efforts will also result in the eventual contribution of name records for all new electronic thesis and dissertation authors to Library of Congress Name Authority Files, which are referenced by libraries as a controlled vocabulary for bibliographic records. Name records will be available through the Library of Congress Linked Data service as URIs. UCSB Open Data and the ADRL content now available as linked data could readily reference these authority headings (Maali 2011).

Different linked data sets do not have to necessarily share a single schema, yet their structure allows them to support cooperation without a need to coordinate. Adopting a linked data approach that defines the relationship between objects regardless of their format, location, or metadata schema, expands the scope of content discovery beyond that which any single system can offer. By extension, this expands discovery beyond an individual campus to the broader research community.

VI. Acknowledgements

We would like to thank the UCSB Library, UCSB Center for Spatial Studies, and Esri Inc. as well as Anabel Ford and Ben Halpern for supporting this research project. The research reported here has been partially supported by an anonymous private donor.

References

1. Addison, A., Moore, J., and Hudson-Vitale, C. (2015). Forging partnerships: Foundations of geospatial data stewardship. Journal of Map and Geography Libraries 11(3): 359–375.

2. Athanasis, N., Kalabokidis, K., Vaitis, M., and Soulakellis, N. (2009). Towards a semantics-based approach in the development of geographic portals. Computers and Geosciences 35(2): 301–308.

3. Avram, H. D. (2003). Machine-readable cataloging (MARC) program. Encyclopedia of library and information science 3: 1712.

4. Baierer, K., Dröge, E., Trkulja, V., Petras, V. (2014). Linked Data Mapping Cultures: An Evaluation of Metadata Usage and Distribution in a Linked Data Environment. In Proceedings of the International Conference on Dublin Core and Metadata Applications.

5. Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. The Future of the Web for Collaborative Science.

6. Brewer, G. D. (2015). The challenges of interdisciplinarity. Policy Sciences 32(4): 327–337.

7. Buckland, M. K. (1997). What is a document?. Journal of the American Society for Information Science (1986-1998), 48(9), 804.

8. Cooley, S., Lafia, S., Medrano, A., Stephens, D., and Kuhn, W. (2015) Spatial Discovery Expert Meeting Final Report. Center for Spatial Studies, University of California, Santa Barbara Library. eScholarship: http://escholarship.org/uc/item/64p820kg.

9. Cragin, M., Palmer, C., Carlson, J., and Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society 4023–4038.

10. Durante, K., and Hardy, D. (2015). Discovery, management, and preservation of geospatial data using hydra. Journal of Map and Geography Libraries 11(2): 123–154.

11. Durante, K., Weimer, K. H. and McGee, M. (2016) "Linked Open Data Modeling for Library Cartographic Resources." Presentation at the Annual Association of American Geographers Conference, San Francisco, CA, March 29.

12. Ford, A., Wernecke, C. (2000). Assessing the Situation at El Pilar: Chronology, Survey, Conservation, and Management Planning for the 21st Century. MesoAmerican Research Center. UC Santa Barbara: MesoAmerican Research Center. eScholarship: http://escholarship.org/uc/item/4qr2x8p3

13. Ford, A. (1995) Archaeological Sites Maya Forest GIS. WWW document, http://discovery.ucsb.opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0

14. Freeston, M. (2004). The Alexandria Digital Library and the Alexandria Digital Earth prototype. In Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries—JCDL 2004, p. 410. New York, New York, USA: ACM Press.

15. Golding, C. (2009). Integrating the disciplines: Successful interdisciplinary subjects. Centre for the Study of Higher Education, University of Melbourne.

16. Goodchild, M. F. (2004). The Alexandria Digital Library Project. D-Lib Magazine, pp. 1–8.

17. Halpern, B. Lester, S. and Grorud-Colvert, K. (2009). PISCO: Partnership for Interdisciplinary Studies of Coastal Oceans. Science of Marine Reserves: Meta-analysis: Global synthesis. KNB Data Repository.

18. Hardy, D., and Durante, K. (2014). A Metadata Schema for Geospatial Resource Discovery Use Cases. Code4lib Journal 25.

19. Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., and Hilf, E. (March 2008). The Access/impact Problem and the Green and Gold Roads to Open Access: An Update. Serials Review 34 (1): 36–40. doi:10.1016/j.serrev.2007.12.005.

20. Heath, T., and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan and Claypool.

21. Hey, T., Tansley, S., and Tolle, K., eds. (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

22. Hu, Y., Janowicz, K., Prasad, S. and Gao, S. (2015-1), Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online. Transactions in GIS, 19: 398–416. doi: 10.1111/tgis.12151

23. Hu, Y., Janowicz, K., Prasad, S., and Gao, S. (2015-2). Enabling Semantic Search and Knowledge Discovery for ArcGIS Online : A Linked-Data-Driven Approach. In AGILE, pp. 1–16.

24. Keßler C, Janowicz, K., and Kauppinen, T. (2012). Exploring the research field of GIScience with linked data. In Xiao N, Kwan M-P, Goodchild M F, and Shekhar S (eds) Geographic Information Science: Seventh International Conference, GIScience

2012, Columbus, OH. September 18–21, 2012, Proceedings. Berlin, Springer Lecture Notes in Computer Science 7478: 102–115.

25. Maali, F., Cyganiak, R., & Peristeras, V. (2011). Re-using Cool URIs: Entity reconciliation against LOD hubs. CEUR Workshop Proceedings, 813.

26. MacMillan, D. (2014). Data Sharing and Discovery: What Librarians Need to Know. The Journal of Academic Librarianship Vol. 40 (5): 541–549.

27. Mayernik, M. S. (2015). Research data and metadata curation as institutional issues. Journal of the Association for Information Science and Technology.

28. Mobility, measured: America is no less socially mobile than it was a generation ago (2014, February 1). The Economist. Retrieved from http://www.economist.com/news/united-states/21595437-america-no-less-socially-mobile-it-was-generation-ago-mobility-measured.

29. Padilla, T. (2016). Humanities Data in the Library: Integrity, Form, Access. D-Lib Magazine, 22 (3/4), 1–12.

30. Patterson, T., Kelso, N. V., & North American Cartographic Information Society. (2009). Natural earth. WWW document https://searchworks.stanford.edu/view/11047527

31. Powell, J. (2014). A Librarian's Guide to Graphs, Data and the Semantic Web. Chandos Publishing. Oxford.

32. Regents of the University of California. "UC Open Access Policies Office of Scholarly Communication." Accessed January 12, 2016. http://osc.universityofcalifornia.edu/open-access-policy.

33. Scaramozzino, J., White, R., Essic, J., Fullington, L. A., Mistry, H., Henley, A., and Olivares, M. (2014). Map Room to Data and GIS Services: Five University Libraries Evolving to Meet Campus Needs and Changing Technologies. Journal of Map and Geography Libraries 10 (1): 6–47.

34. Scheider, S., Degbelo, A., Kuhn, W., and Przibytzin, H. (2014). Content and context—How linked spatio-temporal data enables novel information services for libraries. GIS.Science (4): 138–149.

35. Szukalski, B. (2015, June 17) "ArcGIS Online Demo: A Very Spatial Update." Spatial Discovery Expert Meeting. Upham Hotel, Santa Barbara, California. Lecture.

36. Tsang, Daniel C. (2015). Academic Librarians & Open Access of Data: Challenges & Opportunities in Research Data Management. UC Irvine: UC Irvine Libraries.

37. van Hooland, Seth and Verborgh, R. (2014). Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet.

Chapter 2. Spatial Discovery of Linked Research Datasets and Documents

Sara Lafia, Werner Kuhn

I. Introduction

When university researchers expose their research data, if at all, they publish through various repositories and follow diverse metadata standards. A subset of researchers' published data have open access licenses, but many do not. Regardless of whether data are made open access or not, they often include persistent URIs. Institutions, such as campus libraries, are simultaneously curating accompanying researcher documents as open access manuscripts with persistent URIs, providing access to journal and conference articles through a single-endpoint search. These documents very often reference research data, but this link between data and documents is typically implicit; making the link explicit is currently a tedious manual process (Ballatore et al. 2016).

This work develops support for integrating research data, spatially and thematically, across collections by generating linked metadata for research objects. This approach unifies data, and the documents that reference them, through an intermediary layer that points users to where they are hosted, such as lab servers or publisher repositories. This solution allows researchers to continue their current data publishing and sharing practices. It also does not require libraries to provide support for hosting research data, though they may elect to do so. This results in a library's ability to curate and expose research object metadata without assuming responsibility for maintaining or hosting objects.

A key contribution of our work is a format-agnostic approach to describe data and publications. Related work (Mota and Medeiros 2011) explored shadow-driven document representation, with techniques for extracting core metadata elements from resources, but did not provide for spatial descriptors. Spatial data is highly heterogeneous and can include formats such as shapefiles, imagery, tabular records, and data with implicit spatial references, such as mentioned place names, or toponyms. Such data can be given an explicit spatial "summary" through assignment of a footprint locating their subject. For example, in the case of data containing toponyms, this occurs with the aid of a gazetteer. Our approach of generating a bounding box for research datasets and encoding them in Well-Known Text is a generic means of capturing and describing their spatial "aboutness" (Kuhn et al. 2014). More complex footprint geometries, such as general polygons or multipolygons, may be required in some cases and are compatible with (but not required by) our approach.

We address the question of how the production of spatially referenced and linked metadata can increase the discoverability of research data and documents held in repositories of any sort. Our approach is characterized by a simple model, producing linked metadata that can be queried thematically and spatially. It makes no assumptions about the hosting of data sets or their openness. As long as a research object has a Unique Resource Identifier (URI) and basic metadata adequately supported by existing tools, the data or document can be made discoverable.

II. Method

Selected researchers at the University of California, Santa Barbara (UCSB), interested in sharing their work with a broader community, were recruited to contribute their research data, which are the focus of this work, through a university library open data portal.

Building upon the method of a previous study (Lafia et al. 2016), we have expanded the description of this research data, spatially and semantically. Key developments in the method include the application and extension of a Spatial Metadata Workflow to describe such research data and their publications, an expansion of adopted vocabularies and competency questions, and the formalization of queries. The stepwise procedure undertaken to achieve the study's results is as follows:

1. Recruit university researchers from various domains across campus for study
2. Study existing Spatial Metadata Workflow as applied to Alexandria Digital Research Library
3. Apply workflow to describe datasets in ArcGIS Online Spatial Discovery group
4. Extend workflow to describe other types of research objects, namely documents
5. Extend existing competency questions from previous study
6. Identify and apply appropriate vocabularies for metadata model
7. Test the extended workflow (apply to datasets and document metadata; test and refine metadata model to produce linked metadata; run queries against linked metadata)
8. Elicit stakeholder feedback through workshops (university researchers, library scientists, tool builders)

A. Recruiting campus researchers

One of the major motivations for this work is to facilitate the discovery of research objects across domains and encourage interdisciplinary collaboration. In soliciting partners for the project, it was important to work with researchers at UCSB who had spatial research data readily amenable to sharing. The contents featured on UCSB Open Data reflect the

diversity of these researchers, who range from an archaeologist to a marine biologist. In the previous study, user personas were developed based on these university researchers, which informed the requirements of the open data site and the subsequent Spatial Metadata Workflow (Lafia et al. 2016). Two primary questions that arose when examining the contributed datasets concerned: 1) whether the datasets had explicitly spatial references, such as bounding boxes or named places associated with them; and 2) if the quality of metadata available for the original datasets would prove sufficient for describing their space, time, and theme. Table 1 summarizes the research contents and their contributors.

Table 1. Selected case study documents, datasets, repositories, and contributors.

Document	Repository	Dataset	Repository	Contributor
Assessing the situation at El Pilar	eScholarship	Maya Forest GIS	UCSB Open Data	Dr. Anabel Ford
Acute effects of removing large fish	eScholarship	Sea Bass counts	UCSB Open Data	Dr. Douglas McCauley
Native plant-soil feedbacks	Zotero	Native plant reestablishment	DataONE	Dr. Stephanie Yelenik
Areas of endemism in the Nearctic	Wiley Online	Arthropod Easy Capture	FigShare	Dr. Katja Seltmann

Researchers who agreed to have their data used in the case studies had already made varying provisions for sharing their research through open access means. However, the datasets were in various states of exposure. For example, datasets such as Dr. McCauley's Sea Bass counts, were published on his lab's server as dynamic feature layers that are

updated daily throughout the season as volunteers contribute to the dataset through citizen science observation efforts using a mobile application. The metadata for the feature layers was minimal as it was obtained from ArcServer. This dataset was not initially exposed through a data portal and thus was not discoverable. Similarly, Dr. Ford's Maya Forest GIS collection was available as open access content on local machines at the UCSB Library, but was not available online as feature services. Unlike Dr. MacCauley's feature layers however, Dr. Ford's collection had already been well-described in ISO 19115 metadata, which was generated in partnership with librarians and geographers in the 1990s. Both contributors' datasets were ingested into ArcGIS Online and exposed as feature services, retaining their original metadata along with the minimal descriptor elements of title and description required by ArcGIS Online. The services were then exposed through the Open Data site and the geometry of the datasets were made discoverable, along with metadata and pointers back to the original data sources.

Conversely, Dr. Yelenik's ecological research datasets were already published to a data repository, DataONE, and came with detailed Darwin Core metadata, including spatial descriptors such as a bounding box and place names. Links to this dataset were added to the open data site, referencing the location in DataONE of the open access dataset via its URI. Similarly, Dr. Seltmann's datasets and query were published to yet another repository, FigShare. In a similar fashion, pointers to the original landing page URI for the dataset were added to the open data site. A general call for research data donations from recent alumni resulted in the inclusion of several other research datasets in the open data site, including the sources used in UCSB Geography graduate Dr. Antonio Medrano's PhD dissertation. This approach to recruiting datasets through individuals proved to be effective but time

consuming, as many researchers already adhere to their discipline's best practices for publishing data but have not traditionally thought to share their research through alternative venues, such as open data sites, and have not often set out to describe their resources spatially.

B. Studying existing Spatial Metadata Workflow

The university library recently developed a Spatial Metadata Update Workflow to produce ISO 19139 metadata to describe resources, such as shapefiles, and ingest their metadata into the Alexandria Digital Research Library. This workflow previously relied on a stepwise procedure to capture core spatial, temporal, and thematic elements using ArcCatalog. The spatial elements include an extent for each resource, represented as a bounding box, which can be assigned manually by a librarian, or generated through place name reconciliation against a gazetteer, such as Esri's World Gazetteer. Named places can also be included as keywords, which can come from the resource title, description, or abstract. These are reconciled against Library of Congress Subject Headings and Named Authority Files. The temporal elements include a document date, which can include month or day and is required, and created or revised dates, which are optional. The thematic elements include topic categories, selected from the controlled set of ISO 19115 terms, along with theme keywords, which are also reconciled against Library of Congress Subject Headings and Authorities.

C. Applying workflow to describe ArcGIS Online datasets

While this workflow had been applied to describe prototypical types of spatial data, such as imagery and feature layers that comprise the university's campus map, the workflow had

not been applied to describe other kinds of research objects, such as research datasets or documents. The datasets contributed by campus researchers represent a diverse array of formats, from static images and tables to dynamic feature services. All of these resources are made available through ArcGIS Online, which also provides metadata creation and editing capabilities. This allowed for the Spatial Metadata Workflow to be applied to describe the heterogeneous datasets in the online interface. ArcGIS Online provides support for a variety of metadata standards, including ISO 19139, and provides validation against an XML schema. Regardless of format, the application of the workflow to each dataset in the Spatial Discovery group resulted in metadata that was updated or generated.

D. Extending workflow to describe documents

In addition to spatial datasets, the open data site also hosts links to related documents that reference research data. ArcGIS Online supports a variety of file formats, including document links, which are simply pointer URLs that reference externally hosted content. Many researchers share documents through open access repositories. In the case of University of California researchers, many choose to share their research with eScholarship,[27] which provides persistent URIs to the resources as PDF files with minimal metadata. These document links can also be described in ArcGIS Online using the metadata creation tools. When applying the Spatial Metadata Workflow to describe documents, it was decided that all descriptors, with the exception of spatial extent, also applied to document links. However, while the documents themselves are not spatially referenced, they are linked

[27] http://escholarship.org/

to spatially referenced datasets. Once applied, all research objects in the Spatial Discovery group are described comparably, adhering to the same standard regardless of native format.

E Extending existing competency questions

As the research objects treated with the Spatial Metadata Workflow were more completely described, the previously defined competency questions (Lafia et al. 2016) were extended along with the metadata model. In addition to allowing users to ask about connections between research objects and about research objects in a particular spatial extent, the extended metadata model allows for questions about people, organizations, places, and themes.

Not only does the Spatial Metadata Workflow capture a bounding box for each dataset, which satisfies the original competency question about spatial extent, but also captures named places mentioned in the author's abstract and provided resource title, which are matched against existing named places in DBPedia. Additionally, key metadata capturing authorship and affiliation provide additional means of viewing the provenance of the data. Importantly, datasets can now be explored both by *place* and by *person*, which are arguably the two fundamental systems by which information is cognitively indexed (Mark 2011).

F. Identifying and applying appropriate vocabularies

Originally, the metadata model took advantage of Dublin Core[28] elements to simply link research publications to research datasets. The motivation for selecting this vocabulary is its wide adoption by libraries. Since this first implementation (Lafia et al. 2016), the model has

[28] http://purl.org/dc/elements/1.1/

been expanded substantially to take advantage of SKOS[29] core and GeoLink[30] ontologies to define appropriate classes and properties to relate research objects to resources. The SKOS vocabulary provides classes for *Concepts* and *Collections*, while the Geolink ontology provided the remainder of classes, including *Documents*, *Datasets*, *Person*, and *Place* as well as properties such as *hasPlace* and *hasAuthor*, relating dataset instances to places and authors.

G. Testing the extended production workflow

Putting the method into practice involved applying the revised production workflow to the contents of the Spatial Discovery group using ArcGIS Online. During this process, we identified missing metadata elements as well as general impediments to applying the production workflow at a larger scale. The production workflow resulted in: 1) spatially described datasets, discoverable through their bounding boxes and spatial search using the UCSB Open Data site; and 2) semantically disambiguated metadata for the datasets and documents that link them and enrich data discovery by providing more context about places, time, themes, and authors, discoverable through a triplestore endpoint.

H. Eliciting researcher feedback

In eliciting feedback from researchers, we are primarily interested in learning about: 1) the kinds of data that are not currently treated by our approach, but are of interest; and 2) potential barriers to adoption of such a workflow, from the perspective of any of the project stakeholders, including the researchers, the university library, or technical partners in

[29] http://www.w3.org/2004/02/skos/core
[30] http://schema.geolink.org/1.0/base/main

industry. This feedback will be elicited by usability testing conducted in collaboration with the project stakeholders.

III. Results

Services administered through the university library, in particular ArcGIS Online, which is a cloud-based spatial data management and visualization platform, support researchers by exposing their data in a geographically referenced form. The resulting open data portal, the UCSB Open Data site, leverages ArcGIS Online. Existing RDF vocabularies are applied to describe the shared research data and their relationships with documents in other repositories.

The following steps summarize a proposed workflow that enables the spatial discovery of datasets as well as their semantic annotation. Each step is described in more detail below.

1. **Share** - Researchers share pointers to their research objects using ArcGIS Online.
2. **Describe** - Librarians describe the research objects by metadata.
3. **Aggregate** - Research object metadata and optionally data are aggregated in UCSB Open Data.
4. **Refine** - Tabular metadata elements are cleaned, described with selected vocabularies, and enriched using reconciliation services.
5. **Triplify** - Vocabularies are applied to transform the tabular metadata to triple statements in an RDF skeleton; URIs are minted for the resources.
6. **Query** - Triples are loaded into a triplestore and explored with SPARQL query language.

A. Sharing research objects

The UCSB Open Data site, linked to UCSB Library's ArcGIS Online instance, exposes research data already shared through ArcGIS Online. Open Data is an extension for ArcGIS Online that allows an organization to expose as open access data a subset of contents shared with groups within an organization. ArcGIS Online Open Data is format and metadata agnostic, and allows for any geographically referenced object to be shared through the Open Data portal. Researchers at UCSB are encouraged and guided to share their datasets and documents with the Spatial Discovery group, managed by the university library and the Center for Spatial Studies. Researchers can manage their own content by uploading links to their datasets or documents using ArcGIS Online, as shown in Figure A1 in the appendix. The contents of the researchers featured in this study include static and dynamic datasets and services, hosted in various locations, including lab servers. In the case of static objects, such as shapefiles or imagery, ArcGIS Online provides a mechanism that exposes the datasets as hosted feature services, such as dynamic WFS.

All datasets are assigned a geographic footprint, which is derived from the region that they are about. Librarians can use a gazetteer to translate named places in researcher abstracts to footprints. Researchers can also share related documents in a similar fashion by providing the URI of the open access article in ArcGIS Online, or simply provide the bi-directional reference links between data sets and documents.

B. Describing research objects

The shared research objects, both datasets and publications, are described in accordance with a Spatial Metadata Workflow using ArcGIS Online's metadata editor. The Spatial Metadata Workflow has been developed by UCSB Data Curation and Maps and Imagery

Library for ingesting contents into the Alexandria Digital Research Library[31]. This metadata creation guide was developed as a best-practices manual for describing core metadata elements. ISO 19139 metadata are produced for the spatial datasets, which include controlled topic categories. The documents shared by researchers that reference the data are described using the researcher's name and ORCID[32] when available. Once fully described, the footprints of all research objects are exposed in UCSB's Open Data site[33] as shown in appendix Figure A2. In the case of feature services, Open Data allows for additional GIS operations on the datasets such as filtering, querying, and spatial analysis within the site environment.

C. Aggregating research objects

The research object metadata are downloaded as tabular data from the ArcGIS Online Spatial Discovery group using Administrator Tools[34] as shown in appendix Figure A3. Each record in the generated table describes a research object while the fields are the selected attributes. The core elements captured in the Spatial Metadata Workflow are fields in the table. The bounding boxes for the datasets are represented in ArcGIS Online as two pairs of coordinates, representing its vertices.

Some resources also include alternative coordinate system descriptions. These are first verified to conform to WGS84 Web Mercator, which is required for display by ArcGIS Online, and then are reformatted as Well-Known Text, concatenated, and standardized using

[31] http://alexandria.ucsb.edu/
[32] http://orcid.org/
[33] http://discovery.ucsb.opendata.arcgis.com/
[34] https://github.com/Esri/ago-admin-wiki/wiki/Tools

Refine, which is a browser-based tool for cleaning, transforming, and extending data with web services (van Hoolen et al. 2014).

D. Refining research object metadata

The tabular metadata are imported into Refine with its RDF extension[35]. The inputs are tabular metadata, which come from the ArcGIS Online relational database. The outputs are triple statements, which capture the metadata in semantics closer to natural language, consisting of subjects, predicates, and objects. The terms to describe subjects and predicates come from the adopted RDF vocabularies; the subjects are instances of classes and the predicates are relations. For example, a record of a dataset is an instance of *geolink:Dataset* class and has predicates such as *geolink:hasPlace*. The associated object can be either a literal string, such as "Guatemala" or a resource, like *DBPedia:Guatemala*. This transformation from relational database to triple statement is illustrated in Figure 1.

Refine is also used to perform named-entity recognition on the resource titles, descriptions, and keywords. Refine with RDF extension is used to reconcile elements of the metadata, including ISO 19115 themes, keywords, and alternative titles, against a DBPedia endpoint. Subjects, extracted from dataset alternative titles using Named Entity Recognition (NER) against DBPedia Spotlight, are reconciled against the Library of Congress authority records Subject Headings endpoint, where matching strings are linked to the closest *SKOS:Concept*. Places, also extracted by NER, from dataset alternative titles, are reconciled against *DBPedia:Places*. An example of named NER, extraction, and reconciliation using DBPedia Spotlight is shown in appendix Figure A4.

[35] http://refine.deri.ie/

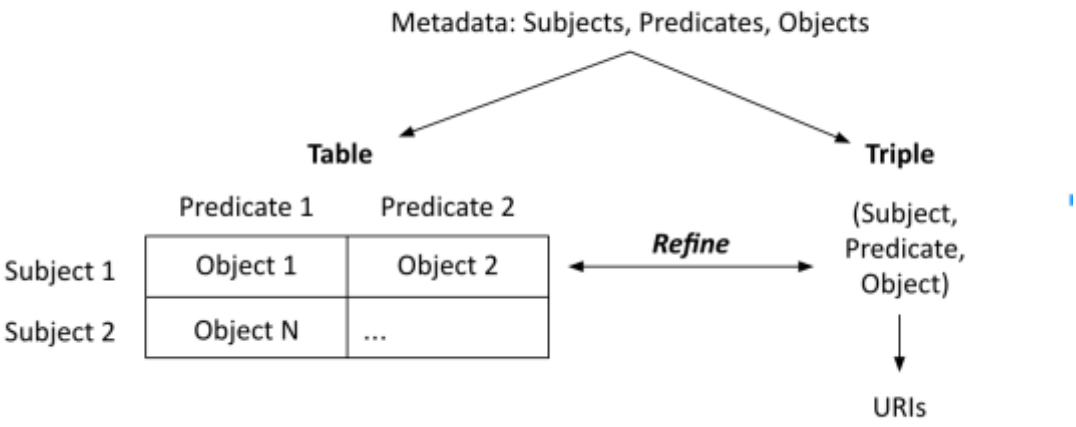


Figure 1. Transforming tabular relational database records into triples.

E. Triplifying research object metadata

Prefixes for Dublin Core (DC), GeoLink (GL), and Simple Knowledge Organization System (SKOS) vocabularies are imported and are applied to the RDF skeleton, shown in Figure 2. The primary node in the triple statement is the dataset, which is described by its URI. The URI is the landing page for the resource in its original hosted location. Secondary nodes are added to the skeleton for Type, Title, Author, Organization, Collection, Year, and Associated Resource. Each dataset is described with the adopted vocabularies. By describing the resources with the Geolink vocabulary, triple statements are generated that enable spatial data querying against existing infrastructure. The Geolink vocabulary provides for geometries, such as bounding boxes, to describe the extent of the resources (Krisnadhi et al. 2015).

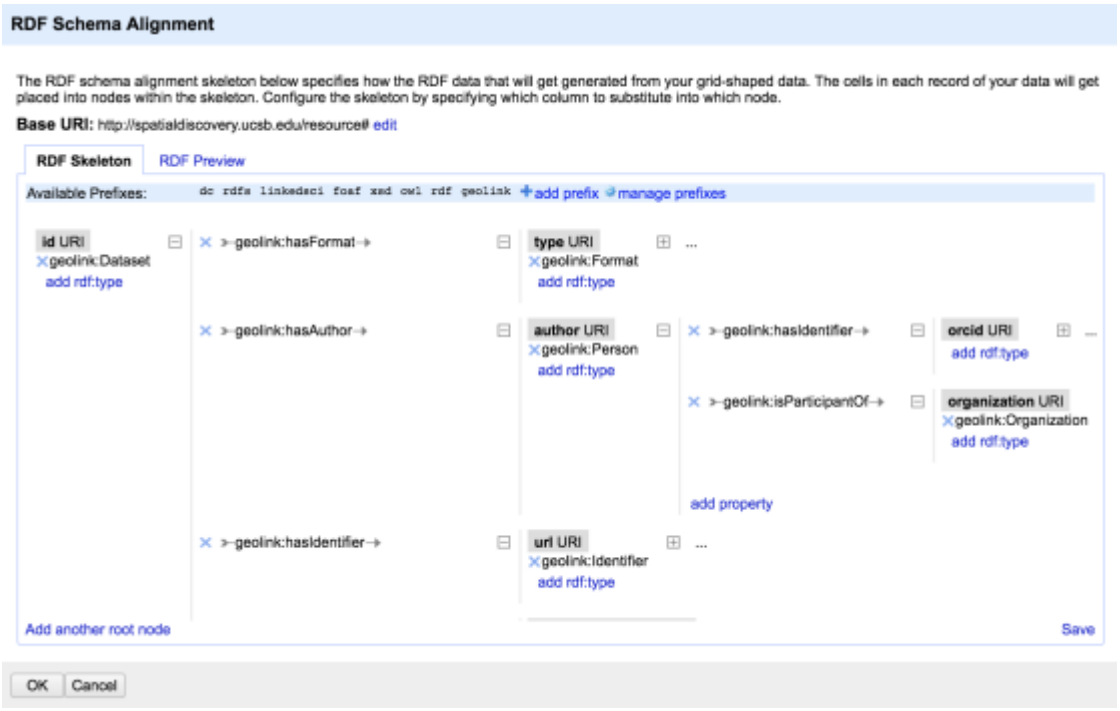


Figure 2. Applying the Geolink ontology to ArcGIS Online dataset metadata.

The CSV columns and rows are transformed into triple statements (subject-predicate-object), based on the imported vocabularies applied to the RDF skeleton. The first step is to mint URIs, which describe the resources. The URIs conform to the standard pattern of authority, container, and item key (Wood et al. 2014). Next, the classes, data properties, and object properties ascribed to each of the metadata fields are aligned against the Geolink, Dublin Core, and SKOS vocabularies, to conform to the desired metadata model, shown in Figure 3. These vocabularies were selected for several reasons. Dublin Core and SKOS are standards currently supported by many academic libraries (Nogueras et al. 2005). Geolink is an ontology developed for describing spatially defined research and supports interoperability with existing web applications (Krisnadhi et al. 2015). The resulting metadata triple statements are serialized and exported as RDF/XML. The RDF (resource description framework) extension to Refine provides a graphical user interface for exporting tabular

data to RDF/XML. The RDF/XML is loaded into a locally built source instance of a Fuseki triplestore, which is shown in appendix Figure A5.

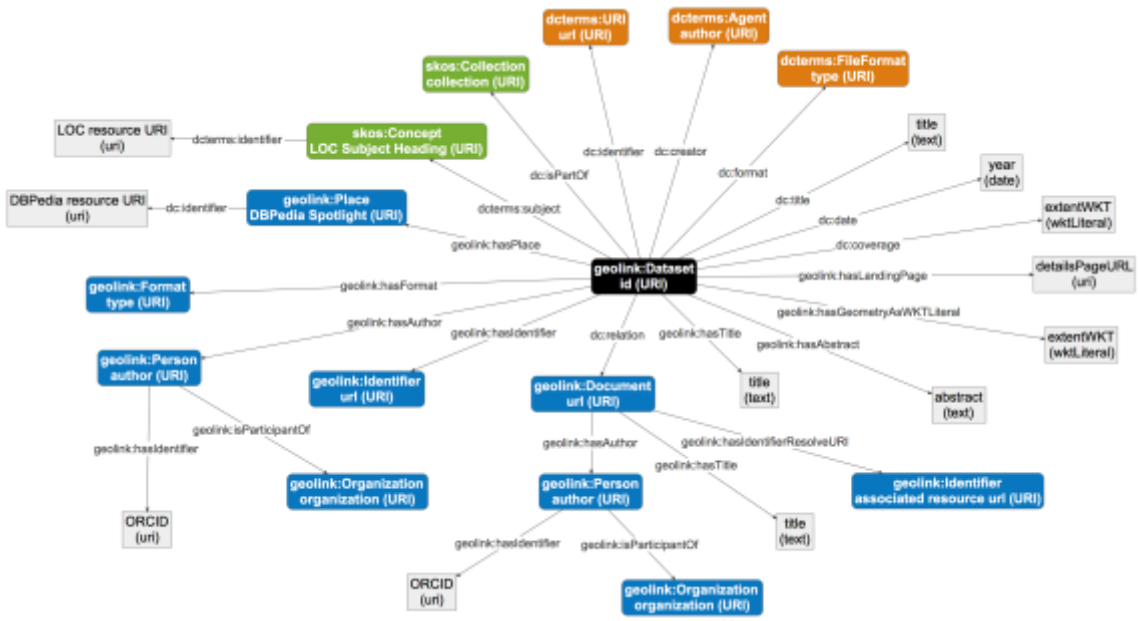


Figure 3. Metadata model adopts Dublin Core, SKOS, and Geolink.

F. Querying research object metadata

The metadata triples are published through a linked data endpoint, which is supported by a backend Fuseki triplestore. Queries can then be run against the endpoint interface. Triple expressions built in the interface are used to query the linked metadata and return matching instances. Alternatively, it is possible to click through object links to discover additional matches along with their associated properties. All of the previously defined relationships captured in the metadata model are now browseable in the linked metadata. Furthermore, resources such as places, themes, and authors are disambiguated, as their URIs provide additional context for understanding what the datasets are 'about' through a *DBPedia:Place*,

a *SKOS:Subject* defined by the Library of Congress, and the author's ORCID, respectively. Figure 4 shows several metadata properties and values for a research object.

property	hasValue
dc:relation	http://escholarship.org/uc/item/0x15m7tq
http://schema.geolink.org/1.0/base/main#hasAbstract	Survey results are available in two seperate formats. The_output contains all non-spatial data from the main survey form, and can be loaded in spreadsheet programs such as Microsoft Excel. The spatial content of the survey is available as a zipped collection of one or more_shapefiles. These files can be opened in GIS applications such as ArcGIS or QGIS. Please note, only completed survey responses are exported. Those still in draft will be excluded. Output columns in both the CSV and shapefile formats are named based on the exportid specified in the form field configuration. If you are looking to analyze spatial data from the shapefiles based on attributes collected in the main response form, you can join fields from the CSV file with spatial features by joining on the_RESPONSE_ID_field.
http://schema.geolink.org/1.0/base/main#hasAuthor	http://spatialdiscovery.ucsb.edu/resource#Douglas+J.+McCauley
http://schema.geolink.org/1.0/base/main#hasFormat	http://spatialdiscovery.ucsb.edu/resource#CSV
http://schema.geolink.org/1.0/base/main#hasGeometryAsWktLiteral	POLYGON(-120.9058 33.1847,117.566 33.1847,117.566 35.1031,-120.9058 35.1031,-120.9058 33.1847) ^^http://www.opengis.net/ont/sf#wktLiteral
http://schema.geolink.org/1.0/base/main#hasIdentifier	http://ucsb.maps.arcgis.com/sharing/rest/content/items/4c3b406e6a9845fea75e292c59ba08f7/data
http://schema.geolink.org/1.0/base/main#hasLandingPage	https://www.arcgis.com/home/item.html?id=4c3b406e6a9845fea75e292c59ba08f7
http://schema.geolink.org/1.0/base/main#hasTitle	Great Giant Sea Bass Count 2014
rdf:type	http://schema.geolink.org/1.0/base/main#Dataset

Figure 4. Instance of dataset metadata annotated with adopted vocabularies.

By defining the types of data that the user would like to retrieve, it is possible to choose a metadata model that meets these requirements. The competency questions identified in the previous study (Lafia et al. 2016) are as follows, and are subsequently translated into SPARQL queries:

- Find datasets referenced by a particular document.
- Find documents that have a particular dataset associated with them.
- Find research objects that overlap with a particular spatial extent.

In addition to discovering data or documents based on a shared link or spatial extent, the updated metadata model allows for more detailed discovery, by person, organization, places and themes. This set of questions is now expanded to enable the following additional queries:

- Find research objects associated with a person (researcher).

- Find research objects affiliated with an organization.
- Explore datasets about places, times, and themes.

```
PREFIX geolink: <http://schema.geolink.org/1.0/base/main#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?dataset ?abstract ?concept ?author
WHERE {
  ?dataset a geolink:Dataset .
  OPTIONAL { ?dataset geolink:hasAbstract ?abstract}
  OPTIONAL { ?dataset dcterms:subject ?concept}
  OPTIONAL { ?dataset dc:creator ?author  }
}
ORDER BY ?concept
```

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX geolink: <http://schema.geolink.org/1.0/base/main#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?dataset ?place ?extent
WHERE {
  ?dataset geolink:hasGeometryAsWktLiteral ?extent ;
           geolink:hasPlace ?place.
}
```

Figure 5. Selected sample SPARQL queries run against Fuseki localhost.

The query structure follows the metadata model by referencing all search by datasets and allowing users to decide which associated links they would like to follow, shown in Figure 5. Exploring the properties of the datasets, which are the central node in the metadata model, also facilitates discovery of linked objects. The resource classes are shown in appendix Figure A6.

IV. Discussion

We demonstrated how to make datasets (shared on ArcGIS Online Open Data) amenable to spatial discovery by describing them with existing RDF vocabularies and producing linked metadata. Spatial search is enabled for datasets, which are linked to documents about them. The metadata triples of both datasets and documents are hosted in an endpoint, which can be

added to a variety of services, including linked gazetteers[36]. This offers researchers a more nuanced means of resource exploration than the traditional keyword search for documents by author or topic (Scheider et al, 2014).

Outstanding questions include how to construct footprints for research objects that are not explicitly spatial. All of the datasets handled in the case studies are explicitly spatial, so providing for their bounding boxes has been relatively easy. It will be valuable to extend this approach to special library collections, such as in architecture or the humanities, that have implicit spatial references to named places, using a gazetteer. This approach will better support spatial search for research objects across collections and disciplines. Additionally, taking advantage of spatial metadata that already conform to GeoSPARQL specifications, such as Well-Known Text, will allow for building spatial queries that leverage semantics (Hart and Dolbear 2013). Expediting the collection of the core elements of datasets and documents, in collaboration with UCSB Data Curators, will allow data contributors to supply core attribute fields that correspond to the metadata model, URIs, and ORCIDs.

Finally, extending the metadata model with additional vocabularies, such as the Linked Science[37] vocabulary, can generate a linked context for the research where the research itself, rather than the researcher or the derived products, are the primary node (Kuhn et al. 2014). Describing resources with this vocabulary will enable explicit connections between researchers and their research. Visualizing the linked open datasets will also enable additional views of the research objects. For example, viewing the results of a query such as *"Show which collections contain resources about lakes published after*

[36] http://adl-gazetteer.geog.ucsb.edu/
[37] http://linkedscience.org/lsc/ns/

2000" as a graph, would provide a deeper understanding of the interconnections and shared properties of attributes in datasets and documents across researchers and domains.

V. Conclusions

Future work will take advantage of GeoSPARQL capabilities of alternative triplestores, including Marmotta, in building queries. Additionally, use of the Alexandria Digital Library Gazetteer as an endpoint for reconciling places rather than DBPedia will connect named places in the university context. Finally, proposing a browser plugin for Open Data will hide the SPARQL query interface from the user, allowing users to perform faceted browsing on research datasets without having to formulate queries in SPARQL syntax.

A vision for researchers at UCSB who would like to make their research objects discoverable includes provision of a streamlined toolkit that assists researchers in the collection of core metadata elements that correspond to the existing metadata model. This toolkit would allow users to create a research context with which they associate research projects and derived objects. A researcher can have multiple projects associated with his or her research context, which is tracked through an identifier, such as an ORCID. The metadata generated by the toolkit point from the research context to these externally hosted objects, and are subsequently published to a publicly accessible endpoint.

The concepts and techniques developed in this article allow users to take multiple views of spatial data and documents, moving from data manipulation in ArcGIS Online, which supports GIS analysis, to data exploration through an endpoint, which supports reasoning. Our research demonstrates a means of streamlined data sharing, document linking, and spatial data discovery. This notion of exposing contents spatially drives interdisciplinary data sharing and integration.

VI. Supplementary Materials

The open data site can be found at UCSB Open Data[38] and the project repository can be found on GitHub at Spatial Discovery[39].

VII. Acknowledgements

We would like to thank the UCSB Library, UCSB Center for Spatial Studies, and Esri Inc. as well as Anabel Ford, Douglas MacCauley, Krzysztof Janowicz, and Katja Seltmann for supporting this research project. The research reported here has been partially supported by an anonymous private donor.

[38] http://discovery.ucsb.opendata.arcgis.com/
[39] https://github.com/saralafia/spatialdiscovery

VIII. Appendix

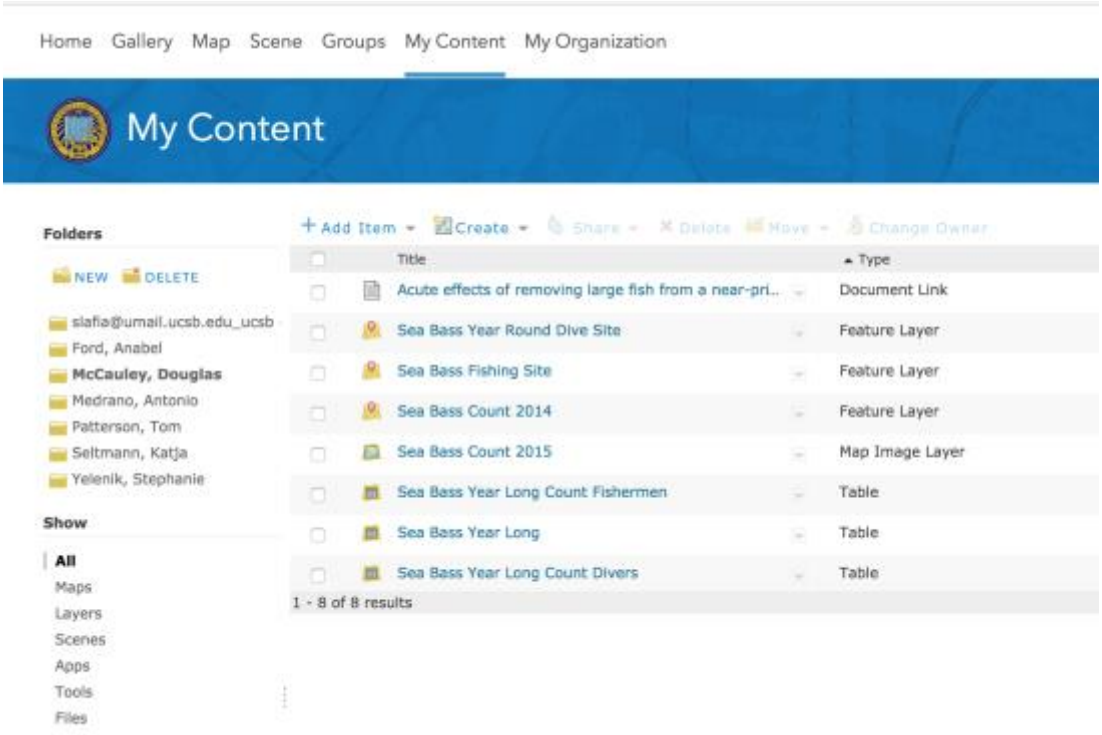


Figure A1. Overview of ArcGIS Online Discovery group content

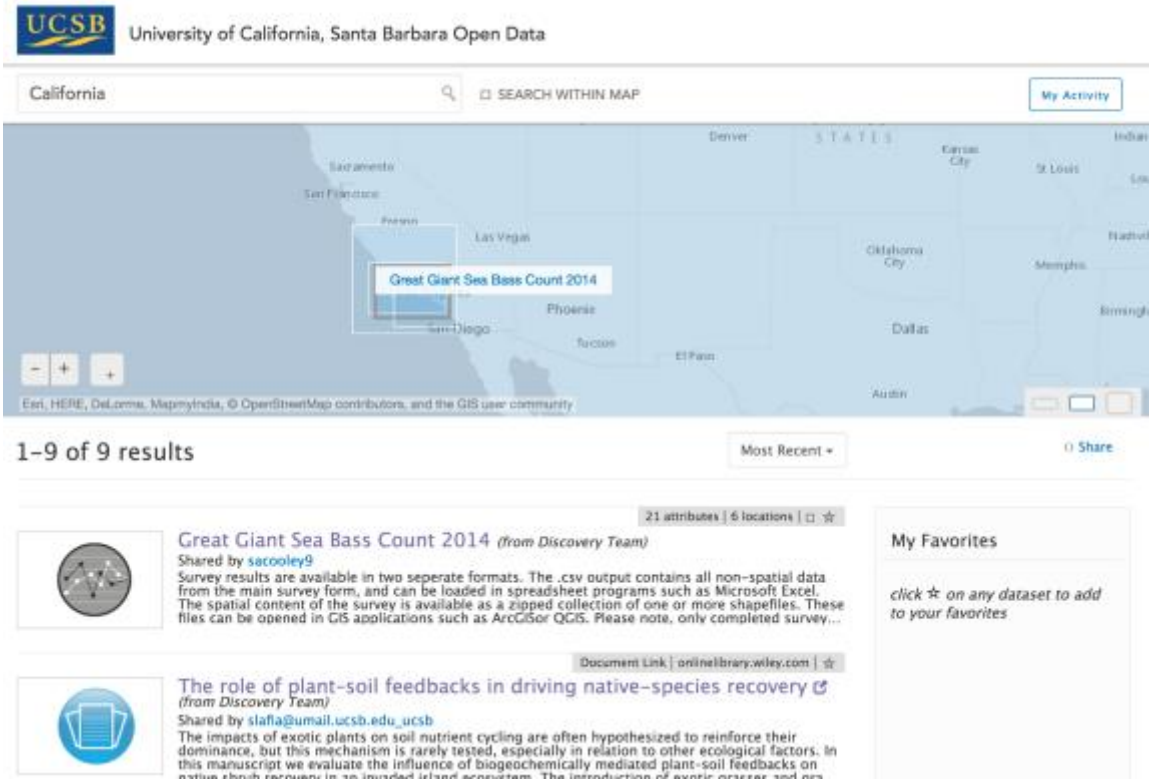


Figure A2. Resulting research object bounding boxes for place "California"

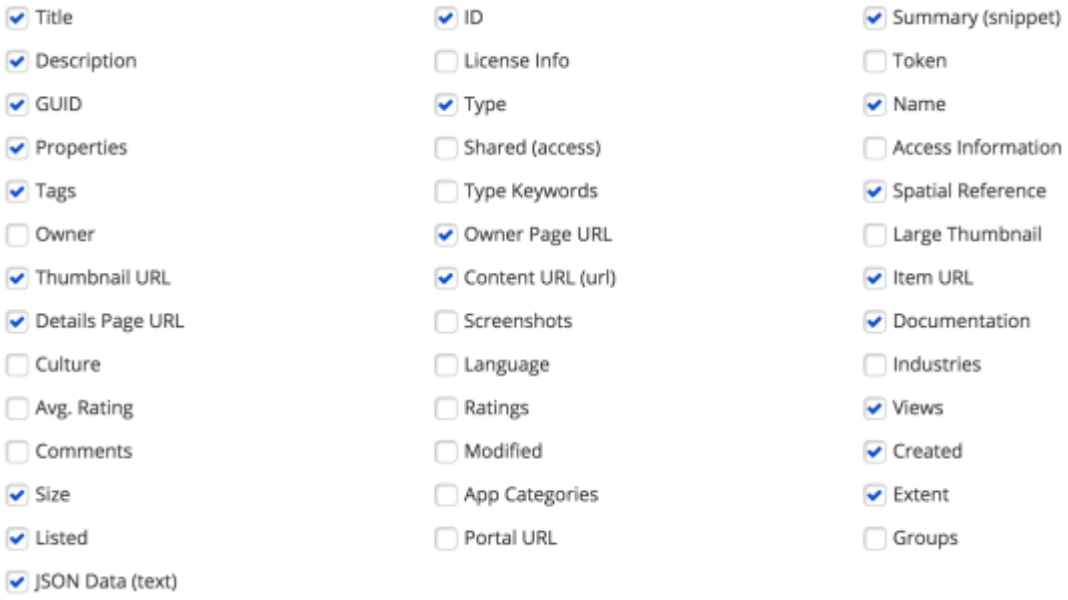


Figure A3. Exported metadata fields from ArcGIS Online Administrator

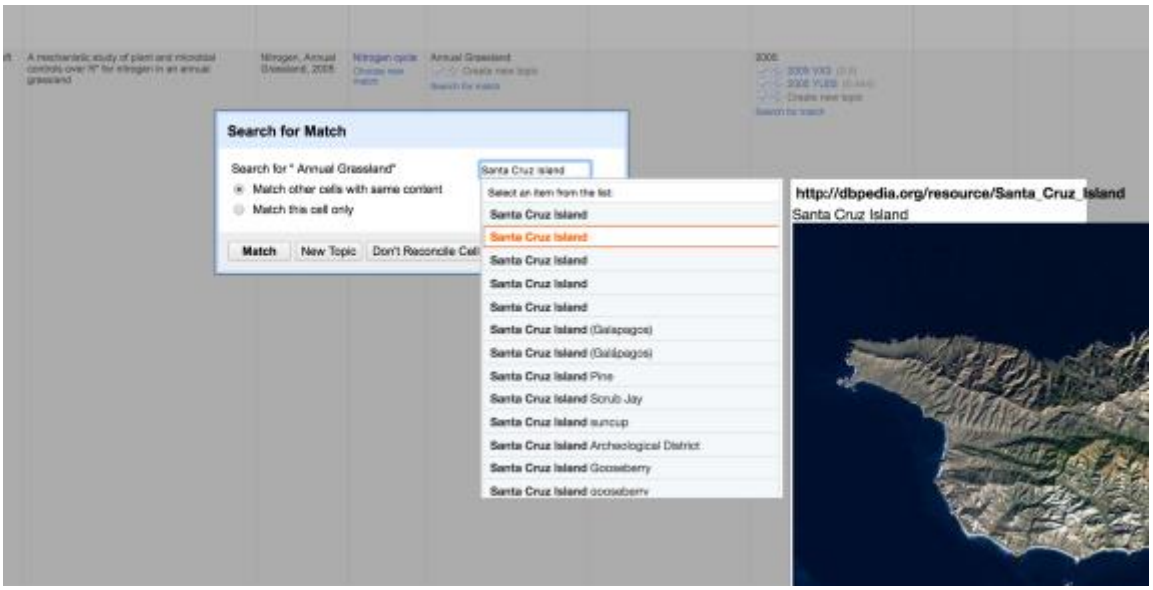


Figure A4. Hosted triples generated from the applied metadata model

Figure A5. Resource classes with prefixes available for query in the triplestore

References

1. Ballatore, F., Kuhn, W., Hegarty, M., Parsons, E. (2016). Spatial approaches to information search. Spatial Cognition and Computation (5868): 1–16.

2. Hart, G.; Dolbear, C. (2013). Linked data: A geographic perspective. CRC Press.

3. Krisnadhi, A., Hu, Y., Janowicz; K., Hitzler, P., Arko, R.; Carbotte, S., Wiebe, P. (2015). The GeoLink modular oceanography ontology. Lecture Notes in Computer Science (9367): 301–309.

4. Kuhn, W., Kauppinen, T., Janowicz, K. (2014). Linked Data - A Paradigm Shift for Geographic Information Science. Springer Lecture Notes in Computer Science (8728): 173–186.

5. Lafia, S., Medrano, A. F., Jablonski, J., Kuhn, W., Cooley, S. (2016). Spatial Discovery and the Research Library. Transactions in (12235): 399–412.

6. Mark, D. M. (2011). Landscape in Language. Transdisciplinary perspectives. Culture and Language Use. (4), 465.

7. Mota, M. S., Medeiros, C. B. (2011). Shadow-driven Document Representation: A summarization-based strategy to represent non-interoperable documents. WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. XI Workshop on Ongoing Thesis and Dissertations.

8. Nogueras-Iso, J., Zarazaga-Soria, F J., Muro-Medrano, P. R. (2005). Geographic information metadata for spatial data infrastructures. Resources, Interoperability and Information Retrieval. Springer.

9. Scheider, S., Degbelo, A., Kuhn, W., Przibytzin, H. (2014). Content and context description - How linked spatio-temporal data enables novel information services for libraries. GIS Science (4): 138–1492.

10. Van Hooland, S., Verborgh, R. (2014). Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet.

11. Wood, D., Zaidman, M., Ruth, L., Hausenblas, M. (2014). Linked Data: Structured Data on the Web. Manning Publications Co.