DOI 10.1111/tgis.12235

ORIGINAL ARTICLE

WILEY Transactions in GIS

Spatial discovery and the research library

Sara Lafia[1] | Jon Jablonski[2] | Werner Kuhn[1] | Savannah Cooley[1] | F. Antonio Medrano[1]

[1] Center for Spatial Studies, 3512 Phelps Hall, Department of Geography, 1832 Ellison Hall, Department of Geography, University of California at Santa Barbara, Santa Barbara, CA 93106-4060, USA

[2] UC Santa Barbara Library, University of California at Santa Barbara, USA

Abstract

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.

KEYWORDS

digital libraries, spatial data portals, Semantic Web

1 | 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 US 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 article.

1.1 | 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, De Roure, Gamble, Goble, & Buchan, 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, Tansley, & Tolle, 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, Lafia, Medrano, Stephens, & Kuhn, 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 Hooland & Verborgh, 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.

1.2 | 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, Palmer, Carlson, & Witt, 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

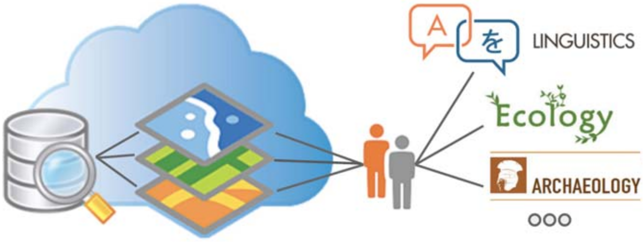


FIGURE 1 Project vision for data discovery and publication integration across domains

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[1], 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 & 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, which has minimal metadata constraints, is semantically linked to a related research object in an institutional environment with tightly controlled metadata.

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.

2 | 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 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.

2.1 | 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).

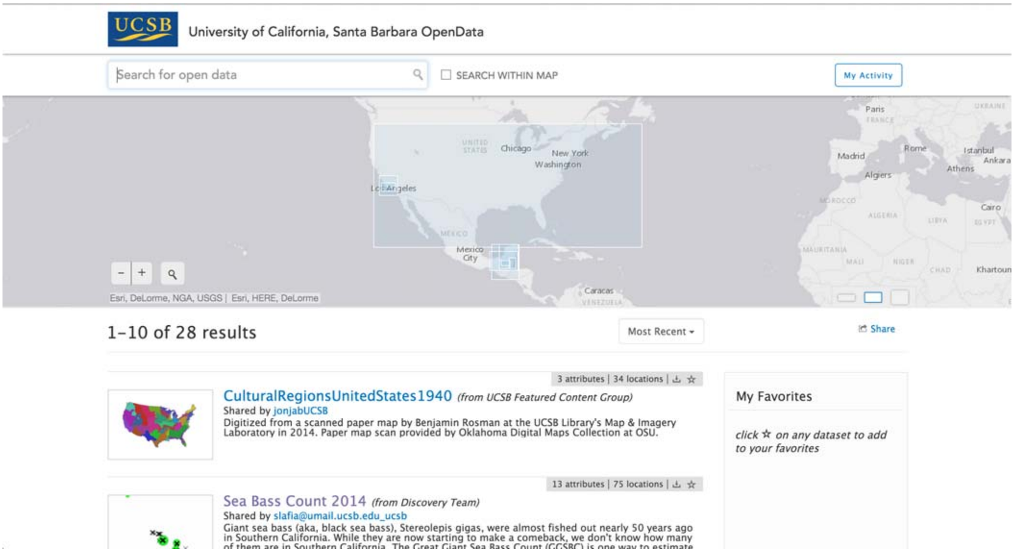
UCSB Map & Imagery Laboratory (MIL) is in the process of developing a spatial metadata workflow using ArcCatalog for the purposes of preparing spatial datasets for ingest into the new Alexandria Digital Research Library (ADRL)[3]. The ISO 19115 standard[4], the Open Geoportal Metadata Creation Guide[5] and the Stanford University metadata creation workflow[6] inform this metadata model. The work described in this article couples these ongoing efforts with the production of linked data.

2.2 | 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, Moore, & Hudson-Vitale, 2015; Durante & Hardy, 2015). It offers users text-based, spatial, and faceted semantic search to enable discovery of GIS-consumable resources across organizations (Hardy & 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, 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[9], 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, 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



1–10 of 28 results

CulturalRegionsUnitedStates1940 *(from UCSB Featured Content Group)*
Shared by jonjabUCSB
Digitized from a scanned paper map by Benjamin Rosman at the UCSB Library's Map & Imagery Laboratory in 2014. Paper map scan provided by Oklahoma Digital Maps Collection at OSU.

Sea Bass Count 2014 *(from Discovery Team)*
Shared by slafia@umail.ucsb.edu_ucsb
Giant sea bass (aka, black sea bass), Stereolepis gigas, were almost fished out nearly 50 years ago in Southern California. While they are now starting to make a comeback, we don't know how many of them are in Southern California. The Great Giant Sea Bass Count (GGSBC) is one way to estimate

My Favorites

click ☆ on any dataset to add to your favorites

FIGURE 2 UCSB's Open Data instance[10] leverages ArcGIS Online

downloading them. Many organizations are adopting Esri's Open Data platform because ArcGIS Online offers web-based analysis and search. UCSB's instance of Esri Open Data[11] is shown in Figure 2.

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[12] (ADRL) repository, which hosts theses and dissertations, or University of California's eScholarship[13], 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.

2.3 | 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 & Verborgh, 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)[14] 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, Weimer, & McGee, 2016). Library of Congress is a notable early organizational contributor to the production of API-accessible linked open data for authority files[15]. Many institutions use these services for authoritative reconciliation (Heath & Bizer, 2011). These services allow institutions, such as the Tate Modern Gallery, to contribute collection metadata to repositories, like GitHub[22], that they neither own nor manage, increasing discoverability, content exposure and creative reuse (Padilla, 2016).

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, Dröge, Trkulja, & Petras, 2014; Hu, Janowicz, Prasad, & Gao, 2015a; Hu, Janowicz, Prasad, &

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-eLinksvia WorldCat
Tom Patterson	political science	World Boundaries of Disputed Areas[20]	EarthWorks	Natural Earth[21]	SearchWorks

Gao, 2015b). 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 & Bizer, 2011). For instance, Athanasis, Kalabokidis, Vaitis, and Soulakellis (2009) described data with domain-specific spatial ontologies in a linked data discovery tool, and Keßler, Janowicz, and Kauppinen (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, Degbelo, Kuhn, and Przibytzin (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.

3 | 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.

3.1 | 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, which are illustrated in Table 1. 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, Lester, & Grorud-Colvert, 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).

The personas 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.

3.2 | 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.

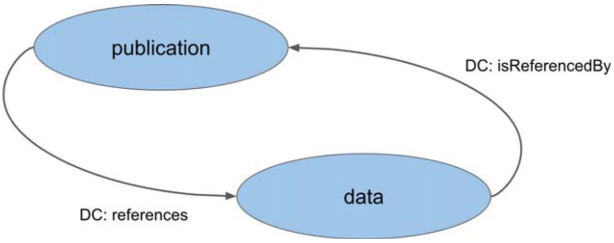


FIGURE 3 Generic Dublin Core Metadata Initiative (DCMI) data model

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.

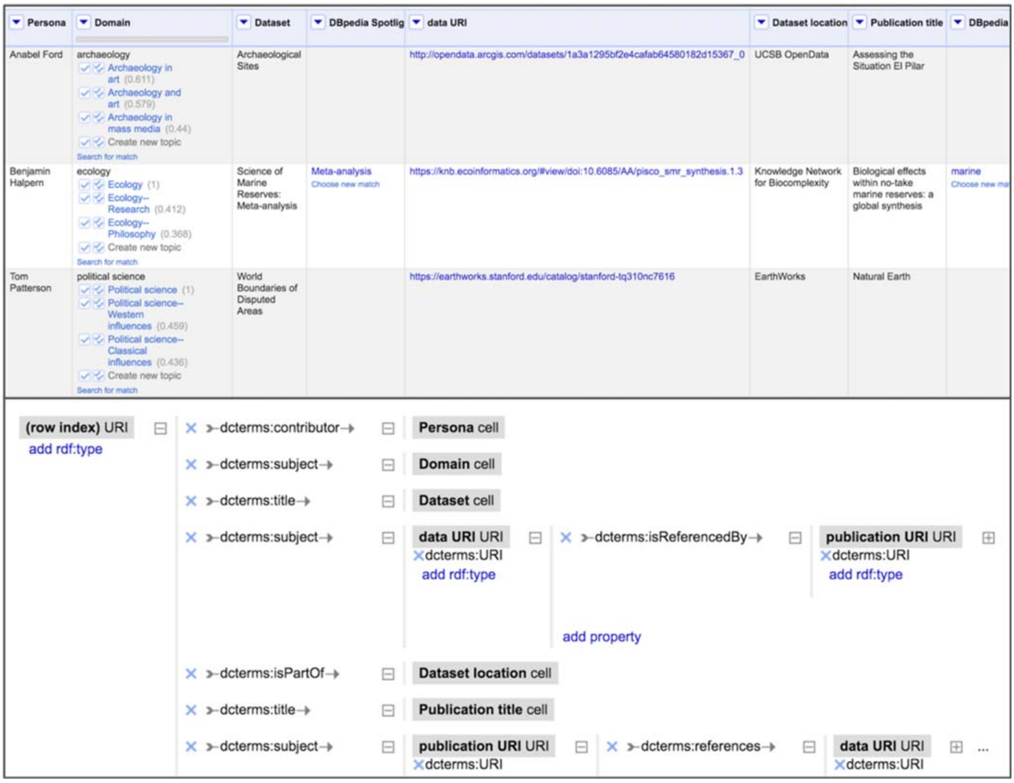
The URIs of research objects correspond to either a data layer or a publication. Datasets that share a common base name are parts of collections and are indicated by URI container. Data creators have only partial control over assignment of the URI resource domain name, which as a best practice, should 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 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[23], shown in Figure 3.

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[24] and even extended to link with other contextually relevant ontologies, such as Wikipedia's knowledge graph DBPedia[25] 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 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.



Persona	Domain	Dataset	DBpedia Spotlig	data URI	Dataset location	Publication title	DBpedia
Anabel Ford	archaeology Archaeology in art (0.611) Archaeology and art (0.579) Archaeology in mass media (0.44) Create new topic Search for match	Archaeological Sites		http://opendata.arcgis.com/datasets/1a3a1295bf2e4cafab64580182d15367_0	UCSB OpenData	Assessing the Situation El Pilar	
Benjamin Halpern	ecology Ecology (1) Ecology--Research (0.412) Ecology--Philosophy (0.368) Create new topic Search for match	Science of Marine Reserves: Meta-analysis	Meta-analysis Choose new match	https://knb.ecoinformatics.org/#view/doi:10.6085/AA/pisco_smr_synthesis.1.3	Knowledge Network for Biocomplexity	Biological effects within no-take marine reserves: a global synthesis	marine Choose new ma
Tom Patterson	political science Political science (1) Political science--Western influences (0.459) Political science--Classical influences (0.436) Create new topic Search for match	World Boundaries of Disputed Areas		https://earthworks.stanford.edu/catalog/stanford-tq310nc7616	EarthWorks	Natural Earth	

(row index) URI
add rdf:type

- × >-dcterms:contributor→ Persona cell
- × >-dcterms:subject→ Domain cell
- × >-dcterms:title→ Dataset cell
- × >-dcterms:subject→ data URI URI ×dcterms:URI add rdf:type
  - × >-dcterms:isReferencedBy→ publication URI URI ×dcterms:URI add rdf:type
  - add property
- × >-dcterms:isPartOf→ Dataset location cell
- × >-dcterms:title→ Publication title cell
- × >-dcterms:subject→ publication URI URI ×dcterms:URI
  - × >-dcterms:references→ data URI URI ×dcterms:URI ...

FIGURE 4 Reconciled OpenRefine template (above) and RDF skeleton (below)

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.

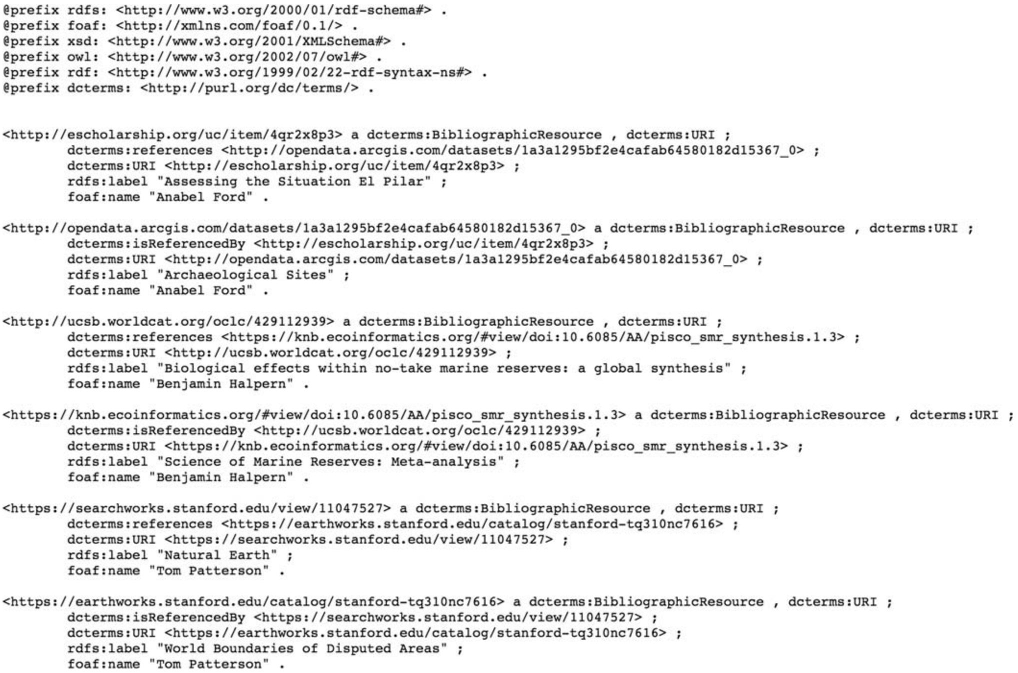
In OpenRefine, a class is a set of RDF resources that use the same templates. Classes such as *publications* and *data* are defined as instances. A new Publications class template uses an RDFa serialization, embedding RDF as triples in HTML documents and encoding the semantic properties and relationships captured in Figure 5.

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.

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 the 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.

4 | 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[26] as the subject (Ford



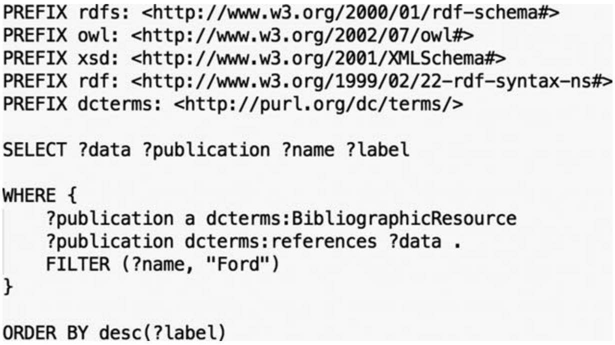
```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/>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


<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" .
```

FIGURE 5 RDF triples for datasets and publications exported in Turtle syntax

and Wernecke, 2000), which are 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.

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



```
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)
```

FIGURE 6 A generic SPARQL query against the triples

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

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.

5 | 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.

5.1 | 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[27] to automate the workflow by running periodic scrapes and generating triplestores from URIs.

5.2 | 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 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 et al., 2011).

Different linked data sets do not necessarily have to 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.

Notes

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

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

[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

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

[9] http://opendata.arcgis.com/

[10] http://discovery.ucsb.opendata.arcgis.com/

[11] http://discovery.ucsb.opendata.arcgis.com/

[12] http://www.alexandria.ucsb.edu/

[13] http://escholarship.org/

[14] https://www.ld4l.org/

[15] http://id.loc.gov/

[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

[22] https://github.com/tategallery/collection

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

[24] http://id.loc.gov/

[25] http://mappings.dbpedia.org/server/ontology/classes/

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

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

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

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

Anon. (2014) Mobility, measured: America is no less socially mobile than it was a generation ago. *The Economist* (2/1/14; retrieved from http://www.economist.com/news/united-states/21595437-america-no-less-socially-mobile-it-was-generation-ago-mobility-measured).

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

Avram, H. D. (2003). Machine-readable cataloging (MARC) program. In M. J. Bates (Ed.), *Encyclopedia of library and information science* (pp. 1712–1730). New York, NY: Marcel Dekker.

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. Paper presented at the International Conference on Dublin Core and Metadata Applications, Sao Paulo, Brazil.

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. Paper presented at the Future of the Web for Collaborative Science Conference, Raleigh, North Carolina.

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

Buckland, M. K. (1997). What is a document? *Journal of the American Society for Information Science*, 48, 804–809.

Cooley, S., Lafia, S., Medrano, A., Stephens, D., & Kuhn, W. (2015). *Spatial discovery expert meeting final report*. Santa Barbara, CA, Center for Spatial Studies, University of California (available at http://escholarship.org/uc/item/64p820kg).

Cragin, M., Palmer, C., Carlson, J., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A*, 368, 4023–4038.

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

Durante, K., Weimer, K. H., & McGee, M. (2016). Linked open data modeling for library cartographic resources. Paper presented at the Annual Meeting of the American Association of Geographers, San Francisco, California.

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

Ford, A., & Wernecke, C. (2000). *Assessing the situation at El Pilar: Chronology, survey, conservation, and management planning for the 21st Century*. Santa Barbara, CA: MesoAmerican Research Center, University of California, Santa Barbara (available at http://escholarship.org/uc/item/4qr2x8p3).

Freeston, M. (2004). The Alexandria Digital Library and the Alexandria Digital Earth prototype. Paper presented at the 2004 Joint ACM/IEEE Conference on Digital Libraries, Tucson, Arizona.

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

Goodchild, M. F. (2004). The Alexandria Digital Library Project. *D-Lib Magazine*, 10(5), 1–8.

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.

Hardy, D., & Durante, K. (2014). A metadata schema for geospatial resource discovery use cases. *Code{4}lib Journal*, 25, 014-07-21.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., . . ., & Hilf, E. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review*, 34, 36–40.

Heath, T., & Bizer, C. (2011). *Linked data: evolving the web into a global data space*. San Francisco, CA: Morgan and Claypool.

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

Hu, Y., Janowicz, K., Prasad, S., & Gao, S. (2015a). Metadata topic harmonization and semantic search for linked-data-driven geoportals: A case study using ArcGIS Online. *Transactions in GIS*, 19, 398–416.

Hu, Y., Janowicz, K., Prasad, S., & Gao, S. (2015b). Enabling semantic search and knowledge discovery for ArcGIS Online: A linked-data-driven approach. Paper presented at the Eighteenth AGILE International Conference on Geographic Information Science, Lisbon, Portugal.

Keßler, C., Janowicz, K., & 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* (pp. 102–115). Berlin, Germany: Springer Lecture Notes in Computer Science Vol. 7478.

Maali, F., Cyganiak, R., & Peristeras, V. (2011). Re-using cool URIs: Entity reconciliation against LOD hubs. *CEUR Workshop Proceedings* 813: 11.

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. *Journal of Academic Librarianship*, 40, 541–549.

Mayernik, M. S. (2015). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67, 973–993.

Padilla, T. (2016). Humanities data in the library: Integrity, form, access. *D-Lib Magazine*, 22, 3/4, 1–12.

Patterson, T., Kelso, N. V., & North American Cartographic Information Society. (2009). Natural Earth. Retrieved from https://searchworks.stanford.edu/view/11047527.

Powell, J. (2014). *A librarian's guide to graphs, data and the semantic web*. Oxford, UK: Chandos Publishing.

Scaramozzino, J., White, R., Essic, J., Fullington, L. A., Mistry, H., Henley, A., & 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, 6–47.

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

Szukalski B 2015 ArcGIS Online demo: A very spatial update. Paper presented at the Spatial Discovery Expert Meeting, Santa Barbara, California.

Tsang, D. C. (2015). *Academic librarians and open access of data: Challenges and opportunities in research data management*. Irvine, CA: University of California, Irvine Libraries.

van Hooland, S., & Verborgh, R. (2014). *Linked data for libraries, archives and museums: How to clean, link and publish your metadata*. London, UK: Facet Publishing.

University of California Regents. (2013). UC Open Access Policies Office of Scholarly Communication. Retrieved from http://osc.universityofcalifornia.edu/open-access-policy.