Spatial Discovery of Linked Research Datasets and Documents at a Spatially Enabled Research Library

Sara Lafia · Werner Kuhn

Sara Lafia
Department of Geography, University of California at Santa Barbara

Werner Kuhn
Department of Geography, University of California at Santa Barbara

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

Abstract

Current publishing practices in academia tend to result in datasets that are difficult to discover. This is because datasets are not well-integrated across academic domains and they are often not linked to the documents that reference them. For these reasons, discovering datasets across domains can be challenging; for example, discovering archaeological observations and biological specimens using the same search is not widely supported, even if both datasets share a similar spatial extent, like Mesoamerica. It is also challenging to retrieve relevant documents that reference datasets; for example, retrieving a series of field reports that reference archaeological observations is typically not supported. Our work develops an extensible method for: 1) geographically integrating collections across disciplinary repositories and 2) connecting datasets to related documents. We describe a collection of spatially-referenced researcher datasets, capturing their metadata elements and encoding them as linked open data. We then leverage existing library services to formalize links from datasets to documents. The system described in this work has been deployed, resulting in an experimental open data site for the UCSB campus. Results indicate that this system can be scaled-up with support from an institutional repository in the near future.

Keywords: digital libraries, spatial data portals, Semantic Web, data integration

1 Introduction

University researchers publish their research datasets through various disciplinary repositories, which follow diverse metadata standards and sometimes provide persistent identifiers. Institutions like university libraries work in parallel to curate researcher documents, like journal articles, as open access manuscripts with persistent identifiers. While documents often cite research datasets, making these links explicit often requires manual effort []. Furthermore, while research datasets may be well-described within disciplinary repositories, such as the KNB (Knowledge Network for Biocomplexity), they are not typically discoverable outside of their academic context, such as ecology. University research data need to be integrated, spatially and thematically, in order to be discoverable, understandable, and ultimately reusable.

In this paper, we address the question of how producing spatially referenced and linked metadata can increase the discoverability of research data and documents held in disciplinary repositories. Our approach produces linked metadata that can be queried thematically and spatially; we do not make assumptions about the hosting of datasets or their openness. As long as datasets or documents have a Unique Resource Identifier (URI) and basic metadata adequately supported by existing tools, they can be made discoverable.

Our approach to describing research datasets and documents, also known as research objects [], is format-agnostic. Related work has explored techniques for extracting core thematic metadata elements from heterogeneous resources, but has not addressed spatial descriptors []. Adopting a format-agnostic approach is key, as spatial data can include take many forms, as shapefiles, imagery, tabular records, and resources with implicit spatial references. For example, if a resource describes a place named "Santa Barbara", this place name can be cross-referenced and disambiguated with the aid of a gazetteer. A footprint can then be assigned to the described object. In this way, we generate and assign generic bounding boxes to research objects.

The experimental open data site resulting from this research exposes data spatially, integrating research across campus by their geographic footprints. We extend the capabilities of this technology through ontology development. Thus, the campus open data site is the result of linking existing services together; this shows that commercial technology already available and widely used on many campuses (e.g. Esri's ArcGIS Online) can be coupled with existing library services and implemented in a university library without additional overhead. While our solution does not insist on openness, it does insist on proven, stable, predictably evolving commercial technology for the API, background mapping, and data visualization engine. By contrast, comparable open-source solutions, like applications that Samvera's technology stack[1], can be fairly opaque and can take teams of developers to customize and implement. We also build upon collaborations with a growing subset of researchers who are already producing geospatial data and are using ArcGIS Online.

To be clear, we do not provide primary hosting for datasets, nor do we make datasets persistent and uniquely identified; this is the role of the repository or archive where data are hosted. Instead, we contribute a method for mapping published datasets into the campus open data site, making them discoverable through geographic footprints and their associated document DOIs. While it is common now to find links to datasets in documents, such as publications, it is less common to find links to such publications from datasets. We are closing this loop through a service that exposes datasets geographically, integrates them thematically, and links them to

[1] http://samvera.org/

document DOIs through their descriptions (to the extent that their authors make such links explicit).

The remainder of this paper is structured as follows. Section 2 contextualizes work related to spatial search in libraries. Section 3 introduces the particular case of the academic library at the University of California, Santa Barbara (UCSB) along with key researcher datasets. Next, in section 4, we expand on the method undertaken to produce linked metadata describing the UCSB researcher datasets and link them to existing documents hosted in a university repository. In section 5, we discuss the implications for this technique in a library context as well as identify key areas for refinement. We conclude in section 6 by illustrating a long range vision for spatial discovery at UCSB.

2 Related Work

Many communities, including geospatial and digital humanities disciplines, recognize the integrative role that location plays in bridging perspectives []. Spatially indexing various types of datasets allows for their integration and discovery in a geographic framework. Existing platforms have developed format-agnostic approaches for data hosting and description, offering a single point of access across hosted services, data types, and thematic domains. For example, the open government data movement has been driven by demands for transparency and has been supported by spatially-enabled platforms such as Esri Open Data[2], CKAN [3], and Socrata[4]. These platforms provide support for data production, dissemination, and reuse across municipal departments, like planning and public works. Work has been done to integrate campus scholarship by adapting such crosscutting open data systems for university contexts [].

Spatially-enabled portals, which span multiple repositories and integrate contents based on their locations, still focus almost exclusively on handling traditional cartographic products []. Platforms, like GeoBlacklight[5], tend to focus specifically on the discovery of traditional map library contents, namely geospatial data. Provisions for the comparable spatial discovery of qualitative data have not yet been made in these systems; textual datasets such as surveys, which typically make reference to location in spatially implicit ways through place name references, are not yet comparably supported.

To address such shortcomings, libraries have utilized linked data models to improve the discoverability of their holdings, including geospatial datasets and text based documents []. Many current efforts, including federal initiatives like BIBFRAME[6] and the Library of Congress Linked Data Service[7], offer authoritative vocabularies and thesauri that can be used to disambiguate both place and theme keywords associated with research objects. In order to semantically enrich metadata, many organizations often turn to these authoritative and well-curated linked data services provided by the Library of Congress []. The benefit of connecting research object keywords to authority files (e.g. linking the thematic keywords "water rights" to a subject heading with the same name) provides both terminological definition and disambiguation, which enables query expansion and inference.

[2] opendata.arcgis.com

[3] http://docs.ckan.org/projects/ckanext-spatial

[4] https://mondara.socrata.com/

[5] https://github.com/geoblacklight

[6] https://www.loc.gov/bibframe/

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

While libraries have long been leaders in improving the discoverability of their holdings, the role of the library in managing repositories for academic research datasets remains unclear. Traditionally, distributed repository frameworks (such as DataONE[8] and GeoLink[9]), oriented around one or more particular disciplines, have provided dataset hosting for researchers who need to ensure persistent access to their research documents and datasets as stipulated by academic grants [].

Today, academic libraries are increasingly called upon to support researchers in meeting these requirements. For example, the University of California (UC) campuses offer preservation and dissemination services for a wide range of scholarship through the eScholarship Repository[10]; researchers can deposit their previously published journal articles to fulfill the UC Open Access Policy. In early 2018, UCSB announced the launch of a campus "Data Collective"[11], which will allow campus faculty and researchers to self-deposit data of all types, with curatorial assistance and overview provided by UCSB Library. Libraries addressing the call to support the curation of campus research are in a unique position to facilitate interdisciplinary data discovery, by providing a domain-neutral meeting ground, and to develop subject-specific guidance for metadata creation, improving the prospects for spatial data discovery.

3 Background

In order to explore faculty data curation needs, we recruited researchers at UCSB who were providing ad-hoc access to their datasets and were motivated to increase the visibility and the discoverability of their research. The Center for Spatial Studies at UCSB offers a Spatial Helpdesk service, which maintains relationships with faculty across campus who work with spatial data. Faculty who had worked with the Spatial Helpdesk in the recent past were invited to participate in a pilot study; respondents were invited to contribute their research datasets for exposure through an open data instance managed by the university library: UCSB's Open Data site[12].

Preparing descriptions of data for exposure involved applying and extending a workflow developed by the UCSB Library[13] to spatially and thematically model the research datasets, along with their related documents []. The Spatial Metadata Update Workflow outlines the basic policies and procedures for updating existing metadata or creating new metadata that describe spatial vector and raster data using Esri ArcCatalog. It includes all elements that are mandatory under the ISO 19115 standard. The workflow provides guidance for metadata creators, namely library staff working with researchers, by standardizing naming conventions for titles (capturing theme and geography), and keywords (linking theme and place to Library of Congress Subject Headings). These standards are informed by the Open Geoportal Metadata Creation Guide[14] and Stanford University's metadata creation workflow.

It is important to note that the case-study developed in this work could likely only be replicated at a research library with a wide array of specialized staff, including those with geospatial expertise. As the scope of this research grows, we also anticipate that subject librarians

[8] https://www.dataone.org/

[9] http://www.geolink.org/index.html

[10] https://escholarship.org/

[11] https://www.library.ucsb.edu/news/faculty-data-collective

[12] http://spatialdiscovery-ucsb.opendata.arcgis.com/

[13] http://guides.library.ucsb.edu/data_management

[14] http://opengeoportal.org/working-groups/metadata/

will play an important role by working directly with researchers to assess data curation needs, including metadata creation, on a case by case basis.

3.1 Recruiting university researchers

One of the major motivations for this work was to support the discovery of academic research across disciplines and encourage interdisciplinary collaboration. For this reason, data were solicited from diverse campus faculty who had research datasets ready to share. Many disciplines are represented in the volunteered data, including archeology, marine biology, ecology, and geography. We assessed each contribution to determine the following: 1) whether the dataset included explicit spatial references, such as bounding boxes or named places; and 2) if the quality of metadata available were sufficient to describe their space, time, and theme. Research data, such as humanities projects with implicit references made to places in literature for example, were not considered due to additional challenges associated with extracting and modeling implicit spatial references. For this reason, we only considered research data that had explicit geospatial references, such as locations defined in geographic coordinates. Table 1 summarizes the data and their faculty contributors.

[Table 1]

Participating researchers had already made varying provisions for sharing their data; thus, the contributed datasets were in various states of exposure. For example, Dr. McCauley, a marine biologist, maintains sea bass counts as published services on his lab's server. They are dynamic feature layers that are updated daily throughout the field season; volunteers contribute to his dataset through citizen science observation efforts using a mobile application. The metadata for the feature layers in McCauley's dataset were minimal. His dataset was not initially exposed through a public data portal and thus was not discoverable. Similarly, archaeologist 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 sets of feature services. Unlike Dr. McCauley however, Dr. Ford had worked with campus librarians to generate ISO 19115 geographic information metadata for her collection.

Contributors' datasets were ingested into ArcGIS Online, which is a collaborative web-based GIS. Datasets were exposed as feature services and they retained their original metadata along with the minimal descriptor elements of title and description required by ArcGIS Online. The services were then exposed through the UCSB 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 had already been published to a data repository, DataONE, and came with detailed Darwin Core metadata, including spatial descriptors such as a bounding box and controlled place names. Links to this dataset were added to the UCSB Open Data site, referencing the location in DataONE of the open access dataset via its URI. Similarly, Dr. Seltmann's datasets and query had been published through another repository, FigShare. Pointers to the original dataset landing page were also added to the UCSB Open Data site.

A general call for research data donations from recent alumni resulted in the inclusion of several other research datasets in the UCSB Open Data site, including 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 adhered to their discipline's best practices for publishing data but had not considered sharing their research through alternative venues, such as open data sites, and in many cases had not yet described their resources spatially.

3.2 Describing datasets online

The datasets contributed by campus researchers represented a diverse array of formats, from static images and tables to dynamic feature services. All of these resources were made available through ArcGIS Online, which also provides metadata creation and editing capabilities. This allowed for the Spatial Metadata Update 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. The workflow was applied to each dataset in the Spatial Discovery group regardless of format, resulting in the update or generation of new metadata.

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[15], 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 Update 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 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.

3.3 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. The method for achieving this and scaling it is described in the following section.

The research objects treated with the Spatial Metadata Update Workflow were more completely described, supporting questions about people, organizations, places, and themes associated with research objects. Not only does the Spatial Metadata Update Workflow capture a bounding box for each dataset, but it also captures named places mentioned in the author's abstract and provided resource title. Additionally, key metadata capturing authorship and affiliation provide additional means of viewing the lineage of the datasets. Importantly, datasets can be explored both by place and by person, which are arguably the two fundamental systems by which information is cognitively indexed [].

[15] http://escholarship.org/

The original metadata model took advantage of Dublin Core[16] elements to simply link research documents to research datasets. The motivation for selecting this vocabulary was its wide adoption by libraries. Since this first implementation, however the metadata model has been expanded substantially to take advantage of other vocabularies such as SKOS[17] core and GeoLink[18] to define appropriate classes and properties to relate research objects to resources. The SKOS vocabulary provided classes for *Concepts* and *Collections*, while the Geolink vocabulary provided classes for *Documents*, *Datasets*, *Person*, and *Place* as well as properties such as *hasPlace* and *hasAuthor*, which relate dataset instances to places and authors. These vocabularies were selected for several reasons. Dublin Core and SKOS are standards currently supported by many academic libraries []. GeoLink is an ontology developed for describing spatially defined research and supports interoperability with existing web applications [].

The decision to expand the metadata model was informed by a need to model more complex relationships among documents and datasets. While the university library largely relies on the flat Portland Common Data Model[19] to describe certain research collections, this model does not provide for the interlinking of objects across collections. Furthermore, the thematic and spatial hierarchies present in research data were not adequately captured. SKOS however provides loose hierarchical relations (such as *broader, narrower,* and *related*) that allow for the association of objects across collections. GeoLink also provides a scientific knowledge base for describing research data, which was absent in the previous data model. The GeoLink vocabulary has already been deployed in participating repositories, allowing for the description of field expeditions, laboratory analyses, and journal publications; its adoption by similar projects made it a suitable choice for describing university research.

4 Method

The UCSB Open Data site leverages ArcGIS Online services, which are administered through the university library and support researchers by exposing their datasets in a geographically referenced form. Existing Resource Description Framework (RDF) vocabularies are applied to describe the shared research datasets and map their relationships with documents held in other repositories. The following steps summarize a workflow developed to semantically annotate research objects and make them spatially discoverable:

1. **Share** - Researchers agree to share their research objects with the university library.
2. **Describe** - Librarians work with researchers to describe their research objects.
3. **Aggregate** - Research object metadata and 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 RDF.
6. **Query** - Triples are loaded into a triplestore and explored with SPARQL query language.

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

[17] http://www.w3.org/2004/02/skos/core

[18] http://schema.geolink.org/1.0/base/main

[19] https://pcdm.org/2016/04/18/models

4.1 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 organizations. 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. The contents 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 dynamic feature services, such as web feature services. Researchers can also share related documents in a similar fashion by providing the identifier of the open access article in ArcGIS Online, or simply provide the links between data sets and documents.

4.2 Describing research objects

The shared research objects, both datasets and documents, are described in accordance with a Spatial Metadata Update Workflow using ArcGIS Online™ s metadata editor. The Spatial Metadata Update Workflow has been developed by UCSB Data Curation and Maps and Imagery Library for ingesting contents into the Alexandria Digital Research Library[20]. 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[21], when available.

All datasets are assigned a geographic footprint, which is derived from the named place that they reference. Librarians can use a gazetteer to look up named places found in textual abstracts and match them to geographic footprints. 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. Once fully described, the footprints of all research objects are exposed in a map interface on UCSB's Open Data site.

4.3 Aggregating research objects

The research object metadata are downloaded as tabular data from the ArcGIS Online Spatial Discovery group using Administrator Tools[22]. Each record in the table describes a research object while the fields are the objects' selected attributes. The core elements captured in the Spatial Metadata Update Workflow are fields in the table. The bounding boxes for the datasets are represented in ArcGIS Online as coordinate pairs, representing their 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. This step is done

[20] http://alexandria.ucsb.edu/

[21] http://orcid.org/

[22] https://github.com/Esri/ago-admin-wiki/wiki/Tools

using Refine[23], which is a browser-based tool for cleaning, transforming, and extending data with web services [].

4.4 Refining research object metadata

The tabular metadata are imported into Refine with its RDF extension. 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 used to describe subjects and predicates come from the adopted RDF vocabularies (SKOS and GeoLink); the subjects are instances of classes and the predicates are relations. For example, a record of a dataset is an instance of the class geolink:Dataset 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.

[Figure 1]

Refine is also used to perform named-entity recognition (NER) on the resource titles, descriptions, and keywords. This is useful because descriptions of research are often replete with spatial and thematic information that can be used to describe research. However, unless these references are made explicit through the creation of links, including named places or themes in research descriptions does not necessarily make them more discoverable. Named Entity Recognition is the first step taken to connect research to other related resources.

In order to check if spatial or thematic descriptions match useful existing web resources that could aid in discovery, Refine with RDF extension is used to reconcile (or look up) elements of the metadata, including ISO 19115 themes, keywords, and alternative titles, against a DBPedia endpoint[24]. DBPedia is an open database of structured and linked concepts derived from Wikipedia. It has been used extensively in research to enrich concepts, build links between concepts, and semantically query relationships, to name several examples [, ,]. Using this technique, we were able to derive representative subjects from the titles of datasets; first, we ran NER using DBPedia Spotlight[25] to identify themes. We then reconciled those themes against Library of Congress authority records Subject Headings[26]. Matching strings are then linked to the closest macthing concept using SKOS:Concept. Similarly, we were also able to extract places from dataset titles using NER in DBPedia Spotlight. We reconciled these against DBPedia:Places.

4.5 Triplifying research object metadata

Prefixes, or abbreviations, for the 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. The GeoLink

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

[24] https://dbpedia.org/sparql

[25] https://github.com/dbpedia-spotlight/

[26] http://freeyourmetadata.org/reconciliation/

vocabulary provides for geometries, such as bounding boxes, to describe the extent of the resources [].

[Figure 2]

This step allows for the transformation of CSV columns and rows into triple statements (subject-predicate-object), based on the vocabularies imported into the RDF skeleton. The first step in this transformation of a flat CSV into linked data triples is to mint URIs describing the resources. The URIs conform to the standard pattern of authority, container, and item key []. 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. The resulting metadata triple statements are serialized and exported as RDF/XML. The RDF extension to Refine provides a graphical user interface, aiding in the transformation of tabular data to triple statements and the resulting export of data to RDF/XML.

[Figure 3]

4.6 Querying research object metadata

Once the triples are exported from Refine as RDF/XML, they are ready to be queried. We load the triples into a locally built server called a Fuseki triplestore. It acts as an endpoint, holding all of the metadata that can be queried. It provides several access protocols, including update and query. The query interface provides two modes of interaction: 1) queries for known properties can be built in the interface using the SPARQL query language, and 2) relationships can be browsed by clicking through links, allowing for the discovery of research objects 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’ in several ways: 1) spatially (using places defined by Wikipedia); 2) thematically (using subjects defined by the Library of Congress); and 3) authoritatively (using the author’s ORCID for tracking), respectively. Figure 4 exemplifies several metadata properties and values for an example research object.

[Figure 4]

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 following competency questions are 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, place, and theme. The initial set of competency questions is now expanded to enable additional queries:

- Explore research objects associated with a researcher.
- Explore research objects associated with an organization.
- Explore research objects by places and themes.

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.

[Figure 5]

5 Results

We demonstrated how to make datasets shared through 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[27]. This offers new methods for exploring campus research repositories beyond the traditional keyword search for documents by author or topic [].

While the set of resources available through the linked data interface will continue to grow, it is already possible to see value added to the discovery of interdisciplinary research. In addition to enabling a geographic view of research data shared through ArcGIS Online, the work described in this paper takes geographic search a step further by linking research data to their creators, to named places, and to other research publications about them. Campus research from diverse departments, ranging from archeology to ecology, have been geographically integrated and have been linked to a variety of disciplinary repositories.

As this research continues, we are interested in eliciting additional feedback from participating researchers in order to better understand: 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 industry. This feedback will be elicited by usability testing conducted in collaboration with the project stakeholders.

In addition to working with stakeholders who are already familiar with Spatial Helpdesk services, we are also broadening our reach. We are partnering with the UCSB Library's Data Curation Program as they launch a campus "Data Collective". This is an opportunity to learn more about spatial data needs and opportunities that exist outside of existing relationships with the Spatial Center at UCSB. The diverse cross-section of disciplines represent sciences, social sciences, and the humanities. While the faculty recruited to participate in this phase of research may already have been concerned with data access and metadata curation before being recruited, the "Data Collective" promises to recruit a more representative sample of research data in various states of curation.

6 Discussion and Conclusions

Outstanding questions include how to construct footprints for research objects that are not explicitly spatial. All of the datasets handled in this study are explicitly spatial, so generating bounding boxes has been straightforward. It will be valuable to extend our approach to special library collections, such as those in architecture or the humanities, that have implicit spatial references to named places. Extending our approach will better support spatial search for research objects across collections and disciplines that are not primarily geospatial. It should be noted that

[27] http://adl-gazetteer.geog.ucsb.edu/

while the creators of research objects (e.g. researchers) may find spatial discovery search advantageous, managers who serve as "meta-users" of collections (e.g. librarians and data curators) and users may find spatial discovery particularly useful for comprehending collections of research objects and the relationships among them. Considering the multifaceted needs of different types of users will improve our approach moving forward.

Additionally, taking advantage of spatial metadata that already conform to GeoSPARQL specifications, such as Well-Known Text, will support spatial queries that leverage geosemantics []. 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.

Finally, extending the metadata model with additional vocabularies, such as the Linked Science[28] 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 []. Describing resources with this vocabulary will enable explicit connections between researchers and their research.

Visualizing the linked open datasets can also enable additional views of the research objects. For example, graphing the results of a query like *Show which collections contain resources about lakes published after 2000*, would provide a deeper understanding of the interconnections and shared properties of attributes in datasets and documents across researchers and disciplines.

While the system developed in this work has not been in use long enough to formally assess the variety of ways in which spatial discovery of research is improving overall, the system has improved the discoverability of research objects by interlinking them. The system 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.

References

[] Ballatore, F., Kuhn, W., Hegarty, M., Parsons, E. Spatial approaches to information search. Spatial Cognition and Computation, 5868, (2016)

[] Bechhofer, S., Gamble, M., Goble, C., Buchan, I., and Roure, D. De. (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nature Proceedings. https://doi.org/10.1038/npre.2010.4626

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

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

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

[] Kuhn, W., Kauppinen, T., Janowicz, K. Linked Data - A Paradigm Shift for Geographic Information Science. Springer Lecture Notes in Computer Science, 8728, (2014)

[] Lafia, S., Medrano, A. F., Jablonski, J., Kuhn, W. Cooley, S. Spatial Discovery and the Research Library, Transactions in GIS, 12235, (2016)

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

[28] http://linkedscience.org/lsc/ns/

[] Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. Journal of the Association for Information Science and Technology, 67(4), https://doi.org/10.1002/asi.23425

[] McGee, M., Durante, K., Weimer, K. H. Toward a Linked Data Model for Describing Cartographic Resources, Journal of Map and Geography Libraries, 13(1), (2017)

[] Mota, M. S., Medeiros, C. B. 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 (2011)

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

[] Scheider, S.; Degbelo, A.; Kuhn, W.; Przibytzin, H. Content and context description - How linked spatio-temporal data enables novel information services for libraries. GIScience, 4, (2014)

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

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

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

Document	**Repository**	**Dataset**	**Repository**	**Contributor**
Assessing the situation at El Pilar	eScholarship	Maya Forest GIS	Open Data	Dr. Anabel Ford
Acute effects of removing large fish	eScholarship	Sea Bass counts	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

Figure 1 Transforming tabular relational database records into triple statements using Refine

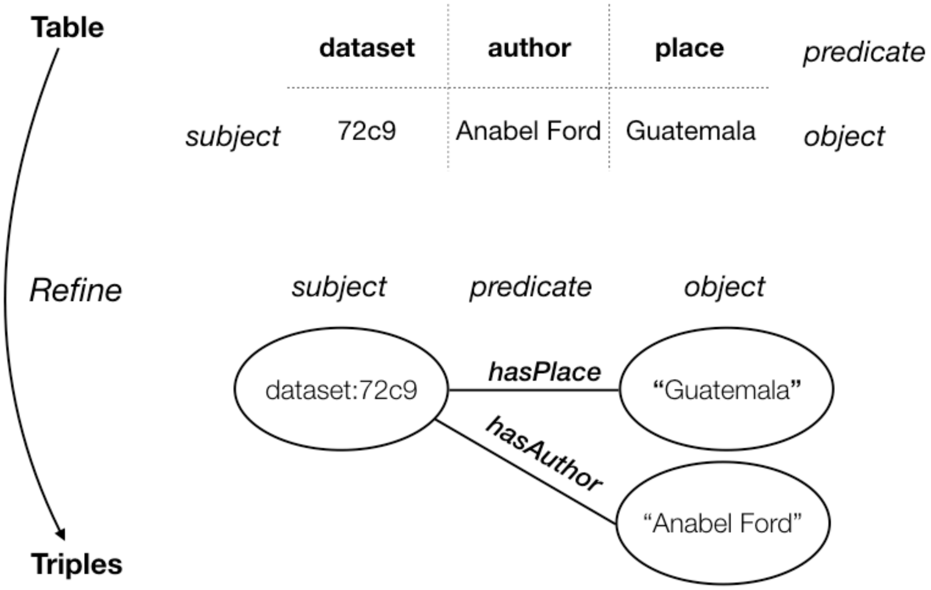
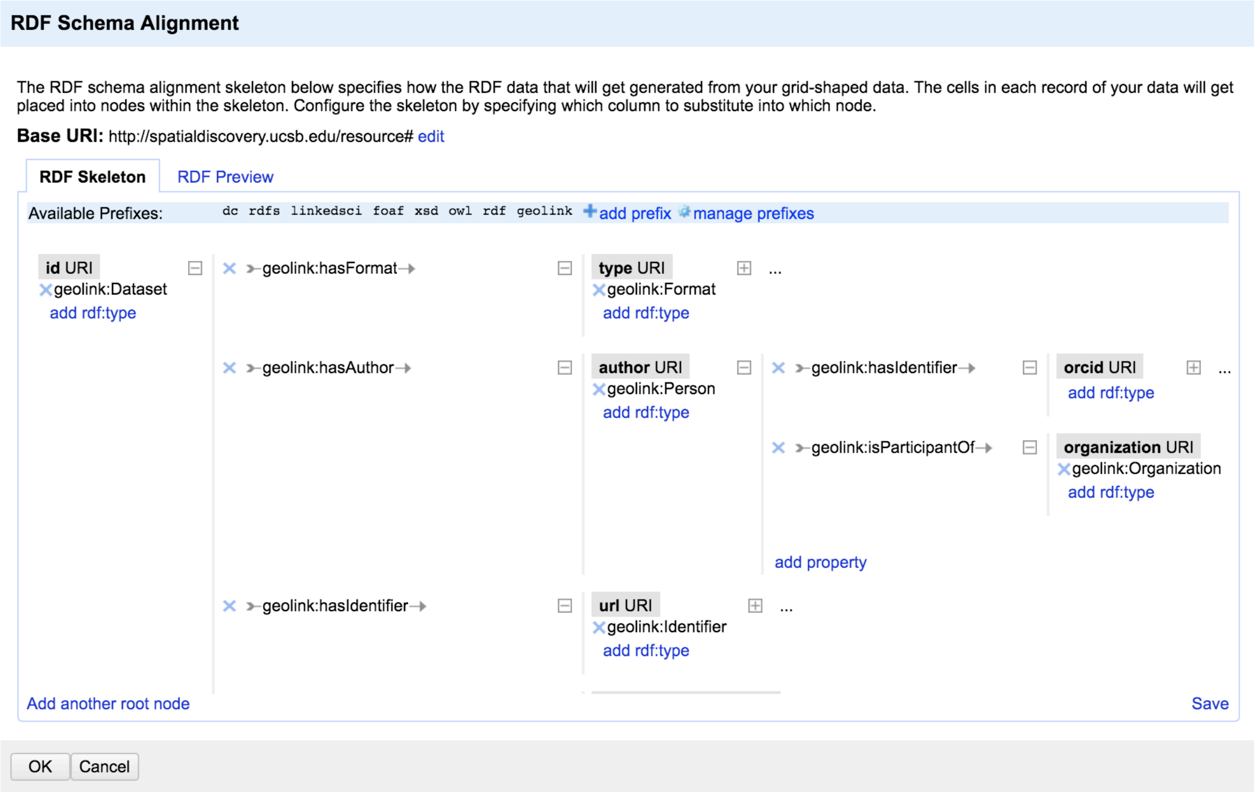


Figure 2 Using Refine with RDF extension to describe ArcGIS Online datasets with the GeoLink vocabulary



RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://spatialdiscovery.ucsb.edu/resource# edit

RDF Skeleton | RDF Preview

Available Prefixes: dc rdfs linkedsci foaf xsd owl rdf geolink + add prefix manage prefixes

- **id** URI — ✕ geolink:Dataset — add rdf:type
  - ✕ geolink:hasFormat → **type** URI — ✕ geolink:Format — add rdf:type ...
  - ✕ geolink:hasAuthor → **author** URI — ✕ geolink:Person — add rdf:type
    - ✕ geolink:hasIdentifier → **orcid** URI — add rdf:type ...
    - ✕ geolink:isParticipantOf → **organization** URI — ✕ geolink:Organization — add rdf:type
    - add property
  - ✕ geolink:hasIdentifier → **url** URI — ✕ geolink:Identifier — add rdf:type ...

Add another root node | Save

OK | Cancel

Figure 3 The metadata model is constructed from adopted vocabularies that include Dublin Core, SKOS, and the Geolink ontology, where *geolink:Dataset* is the primary node resource

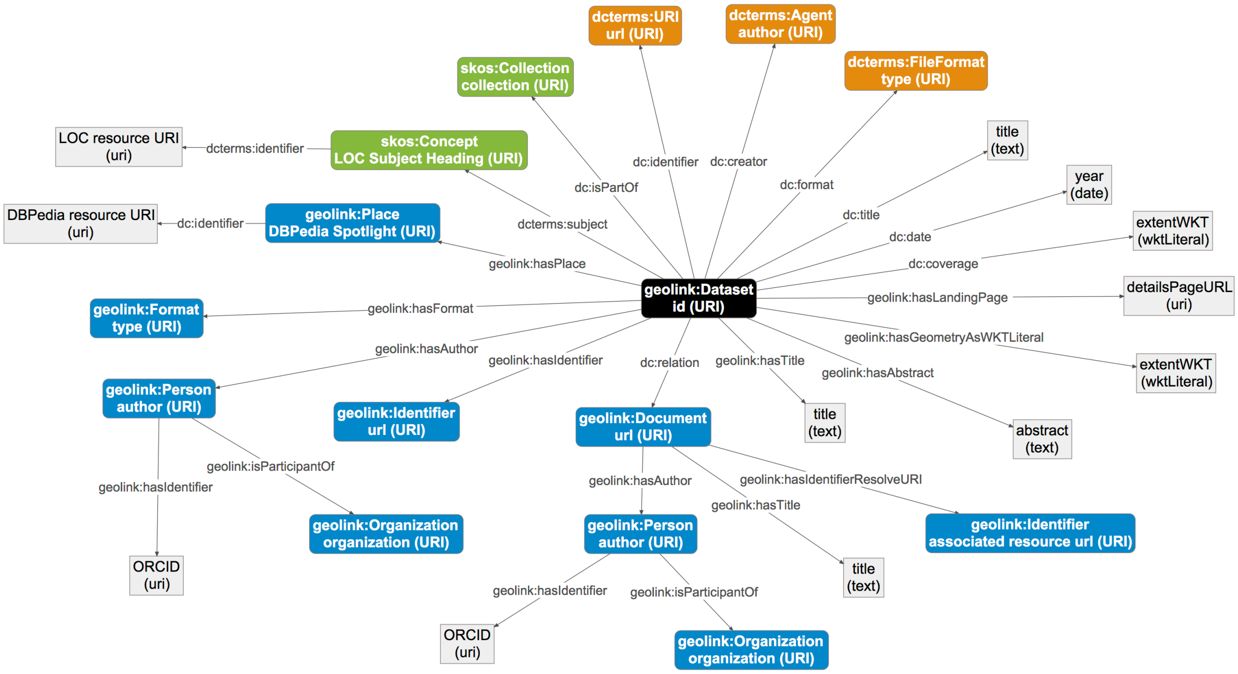
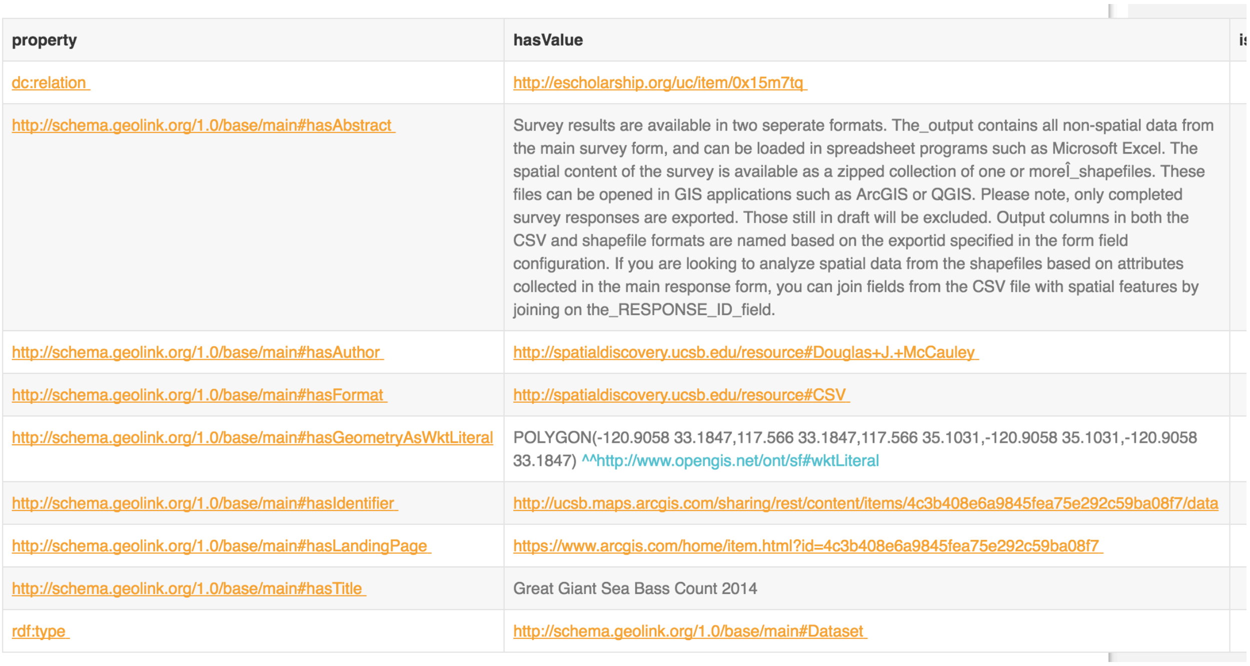
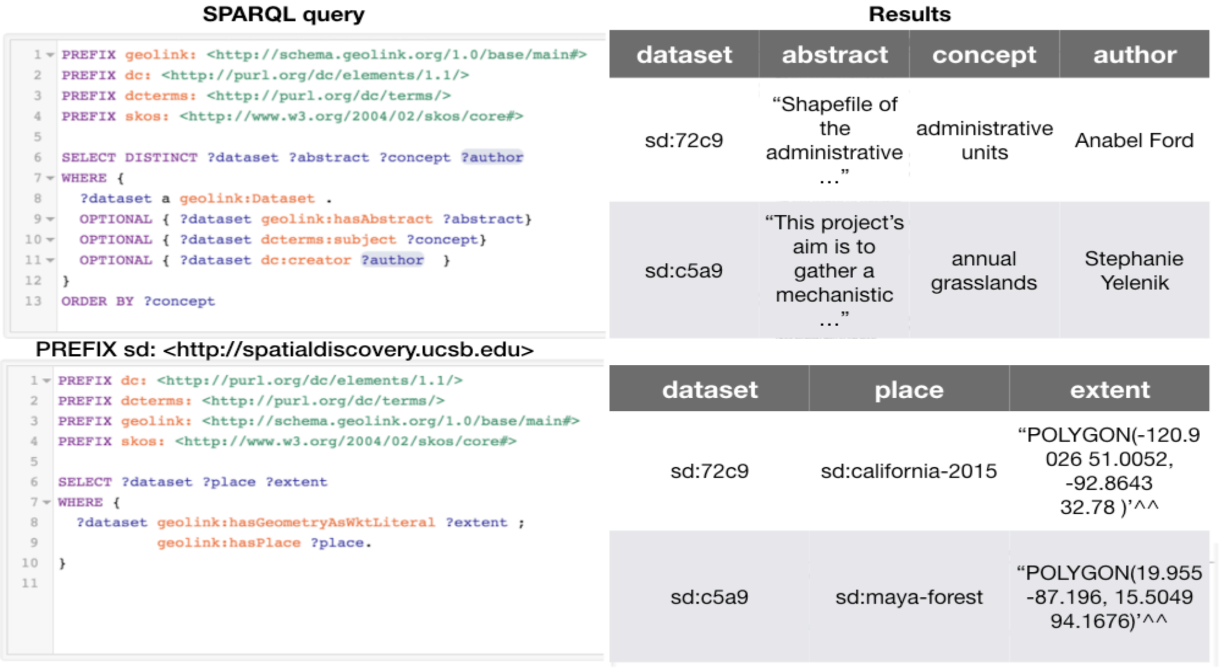


Figure 4 An instance of a dataset's metadata annotated with the adopted vocabularies



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 moreÎ_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/4c3b408e6a9845fea75e292c59ba08f7/data
http://schema.geolink.org/1.0/base/main#hasLandingPage	https://www.arcgis.com/home/item.html?id=4c3b408e6a9845fea75e292c59ba08f7
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 5 Selected sample SPARQL queries (left) run against Fuseki localhost with results (right) for (**a**) Themes, abstracts, and authors of datasets (**b**) Extent and place of datasets



SPARQL query

```
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
```

Results

dataset	abstract	concept	author
sd:72c9	"Shapefile of the administrative …"	administrative units	Anabel Ford
sd:c5a9	"This project's aim is to gather a mechanistic …"	annual grasslands	Stephanie Yelenik

PREFIX sd: <http://spatialdiscovery.ucsb.edu>

```
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.
}
```

dataset	place	extent
sd:72c9	sd:california-2015	"POLYGON(-120.9026 51.0052, -92.8643 32.78)'^^
sd:c5a9	sd:maya-forest	"POLYGON(19.955 -87.196, 15.5049 94.1676)'^^