

# UC Santa Barbara

## UC Santa Barbara Previously Published Works

### Title

Improving Discovery of Open Civic Data

### Permalink

<https://escholarship.org/uc/item/6rn1b22w>

### ISBN

978-3-95977-083-5

### Authors

Lafia, Sara  
Turner, Andrew  
Kuhn, Werner

### Publication Date

2018-02-08

### DOI

10.4230/LIPIcs.GISCIENCE.2018.9

### Data Availability


The data associated with this publication are available at:  
<https://github.com/saralafia/esri-hub>

Peer reviewed

# Improving Discovery of Open Civic Data

**Sara Lafia**

Department of Geography, University of California, Santa Barbara, USA  
slafia@geog.ucsb.edu


 <https://orcid.org/0000-0002-5896-7295>

**Andrew Turner**

Esri DC, Office of Research and Development, Arlington, VA, USA  
ATurner@esri.com

**Werner Kuhn**

Department of Geography, University of California, Santa Barbara, USA  
werner@ucsb.edu

 <https://orcid.org/0000-0002-4491-0132>

---

## Abstract

We describe a method and system design for improved data discovery in an integrated network of open geospatial data that supports collaborative policy development between governments and local constituents. Metadata about civic data (such as thematic categories, user-generated tags, geo-references, or attribute schemata) primarily rely on technical vocabularies that reflect scientific or organizational hierarchies. By contrast, public consumers of data often search for information using colloquial terminology that does not align with official metadata vocabularies. For example, citizens searching for data about bicycle collisions in an area are unlikely to use the search terms with which organizations like Departments of Transportation describe relevant data. Users may also search with broad terms, such as “traffic safety”, and will then not discover data tagged with narrower official terms, such as “vehicular crash”. This mismatch raises the question of how to bridge the users’ ways of talking and searching with the language of technical metadata. In similar situations, it has been beneficial to augment official metadata with semantic annotations that expand the discoverability and relevance recommendations of data, supporting more inclusive access. Adopting this strategy, we develop a method for automated semantic annotation, which aggregates similar thematic and geographic information. A novelty of our approach is the development and application of a crosscutting base vocabulary that supports the description of geospatial themes. The resulting annotation method is integrated into a novel open access collaboration platform (Esri’s ArcGIS Hub) that supports public dissemination of civic data and is in use by thousands of government agencies. Our semantic annotation method improves data discovery for users across organizational repositories and has the potential to facilitate the coordination of community and organizational work, improving the transparency and efficacy of government policies.

**2012 ACM Subject Classification** Information systems → Digital libraries and archives

**Keywords and phrases** data discovery, metadata, query expansion, interoperability

**Digital Object Identifier** 10.4230/LIPICs.GIScience.2018.9

**Supplement Material** <https://github.com/saralafia/esri-hub>

**Acknowledgements** The work presented in this paper was part of a research internship of the first author at Esri’s Research and Development Office. Additional contributions by Pranav Kulkarni, Daniel Fenton, and Alexander Harris of Esri Research and Development are gratefully acknowledged. The work was supported by Esri and by UCSB’s Center for Spatial Studies.



© Sara Lafia, Andrew Turner, and Werner Kuhn;  
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 9; pp. 9:1–9:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

In recent decades, great strides have been made to encourage data creators and providers to make the findings of their research or results of their activities publicly accessible. Researchers receiving grant funding now face mandates to preserve and expose data resulting from their research [9]. Parallels can be drawn between the mounting movement surrounding open access in academia and similar movements well underway in the civic arena surrounding shared municipal data; all levels of government, from Federal agencies to city governments, have started exposing data [14]. Open data, also known as open Public Sector Information, contribute to citizens' rights to public access of government information. Open data policies at various levels of government have stimulated and guided the publication of both spatial and non-spatial government data [15]. The resulting creative downstream use of civic datasets is staggering, ranging from mobilization of grassroots citizen initiatives to uptake by private application developers [7]. By making civic data about a range of topics, from departmental budgets to bicycle collisions, consumable through APIs, governments such as the City of Los Angeles<sup>1</sup> have become better connected to their citizenry.

However, simply making data accessible online does not guarantee their discoverability [1]. The likelihood of discovering thematically relevant geospatial data is still quite low; this is due to two key geospatial issues. The first issue is that data produced by co-located and adjacent governments are often described differently. Thus, discovering spatial data about bicycle collisions provided by neighboring governments, such as Arlington and Fairfax, VA along with state data, for example, is not trivial. This is because data, such as bicycle collision statistics, are described in a heterogeneous way by neighboring municipalities and by various levels of government. A second issue is that civic data are not described using terms that public consumers use. Governments may collect and provide traffic collision statistics, while consumers may want to assess community safety for cyclists.

It is unrealistic to imagine all providers of civic data conforming to a single metadata standard or providing suites of additional colloquial keywords to resolve these issues. In fact, the multiplicity of inward-looking open data policies at various levels of government make this untenable [15]. Instead, we ask how semantic mappings can bridge the gap between terms used in peoples' daily lives and terms from technical governmental metadata, thus improving the recall and precision of open civic data. Our approach bridges data provider and data user terms by developing a crosscutting base vocabulary that expands core geospatial themes and can be used to better describe civic data. We demonstrate the value of our approach by applying the vocabulary to automatically annotate data on a novel open access platform.

The contributions of this work are as follows:

- A **system** for harvesting provider-contributed data descriptions
- A **base vocabulary** of core geospatial themes mapping provider to consumer descriptions
- A **protocol** for semantically annotating data with core geospatial themes for consumers

The remainder of this paper is organized as follows. Section 2 provides background on the studied open data platform. Section 3 surveys challenges of and approaches focused on improving data discovery. Section 4 discusses the method developed to enrich tags during metadata harvesting. Section 5 describes the resulting implementation. Section 6 discusses the results of the work and presents a research outlook.

---

<sup>1</sup> <http://geohub.lacity.org/>

## 2 Background

In order to validate the method design and evaluate results, this work integrates semantic annotation into an open access collaboration platform, Esri's ArcGIS Hub<sup>2</sup>. This platform exposes organizational data via ArcGIS, which is a geospatial data management, visualization, and analytics system used by governments, industry, academia, and other organizations to support planning and operations. ArcGIS integrates desktop software with cloud-hosted tools and data services for distributed information access that can be shared privately or with the public. Using ArcGIS Online, members of organizations and the public can create, edit, and share maps and other data. This global system organizes a content-rich catalog of information across a breadth of scientific themes and operational domains.

ArcGIS Hub is a new open access platform that supports and organizes civic engagement and direct collaboration between governments and their constituents. ArcGIS Hub extends the ArcGIS Online system with new capabilities for open data sharing, configurable metadata catalogs, integration with regional and national metadata registries such as Data.gov<sup>3</sup>, and analysis tools for the public to visualize and share perspectives on data relationships.

Governments and other enterprises can use ArcGIS Hub to create custom websites for open data sharing that allow the public to easily search, access and download data. ArcGIS Hub's primary audience are the general public: people and groups outside of the organizations sharing the data. While ArcGIS Hub integrates with proprietary software, it also serves as a standalone platform that enables anonymous, public access to datasets from any other platform or data provider; it is not necessary to have any authentication credentials in order to discover or use open access data shared through ArcGIS Hub.

As of early 2018, over 100,000 datasets had been made available through ArcGIS Hub by more than 5,000 governments, academic institutions, and other organizations. These datasets are discoverable by search term, specified by user keyword, and by area of interest, which can be specified by map interface. The current state of search in ArcGIS Hub is based on keyword matching, which matches user queries against dataset titles, descriptions, and tags. A limitation of this type of search however, is that it fails to capture broader or related contexts of the query, only returning content that has a title, description, or tags containing the input term. For example, a search for "bicycle" would not return related content, such as "pedestrian", or broader content, such as "transport".

Civic data providers are primarily focused on making their data available and secondarily focused on making their data discoverable to public consumers, often only providing descriptions or tags when required and often using domain-specific terms. This creates semantic and schematic barriers to data discovery, resulting in a gulf between terms that users and terms that providers use to describe and search for the same data. Resulting challenges to discoverability and current approaches to address them are the focus of the next section.

## 3 Challenges and Approaches

Data shared through public repositories satisfy basic accessibility requirements, but are often siloed and difficult to discover. A recent report from the Open Research Data Task Force [13] found that the two main challenges to using open data are: 1) finding data to use and 2) (re)using them. While this is especially true of academic data scattered across diverse

---

<sup>2</sup> <http://hub.arcgis.com/>

<sup>3</sup> <https://www.data.gov/>

domain repositories, it is also true of civic data. The current silos for civic data are not simply organizational, but semantic and schematic, rooted in the technical vocabularies used to categorize and structure data [4]. The main challenges to reusing civic data are the domain-specific terms used to describe data and their attribute schemata [13].

Innovations from the arenas of academia, government, and industry demonstrate contrasting, yet complementary, approaches to addressing discoverability challenges [9]; advances in each arena also inform this work. Recent innovations in discoverability have resulted from the implementation of linked data technologies, which allow for data to be self-describing [2]. The uptake of linked data technologies has resulted in an ever-expanding graph of shared knowledge<sup>4</sup>, replete with reusable ontologies from many domains. Linked data technologies address key semantic and schematic challenges, aiding in many arenas such as in the discovery of scientific data for reuse and discovery across integrated civic data streams [1, 11].

### 3.1 Semantic Challenges

The first challenge to civic data discovery is semantic. Semantic heterogeneity is understood to result from differing mental models of phenomena as well as from differences in naming conventions; naming heterogeneity can be overcome with term mappings using thesauri, but cognitive heterogeneity is understood to be a more difficult problem to solve in the absence of a minimum set of common definitions [5]. Our work focuses on overcoming heterogeneous naming of semantically similar content, resulting from divergent metadata standards.

The rigor and quality of data classification and tagging schemes can vary greatly by data provider. In the case of highly curated data, such as Federal data layers shared through Esri's Living Atlas of the World,<sup>5</sup> tags for each dataset have high agreement and control, grouping the data into one of several predefined themes: demographics, transportation, landscape, oceans. . . Similarly, data conforming to the ISO 19115 metadata standard<sup>6</sup> adhere to a highly controlled vocabulary describing what the contents are about by keyword: agriculture, biota, economy, health. . . However, as of early 2018, only 66,000 (about 8 percent) of the 760,000 items in the ArcGIS Hub catalog had formal metadata.

Metadata files in ArcGIS Hub are also not indexed for search; instead, keyword search in Esri's ArcGIS Hub is based on search by regular expression against the titles, descriptions, and tags of content. Organizations contributing data supply their own tags and descriptions, which results in varying levels of quality. Relatively few tags are based on a controlled vocabulary and descriptions of data have varying levels of completeness. This results in a situation where search for "bicycle collisions" returns results for Washington D.C. where data have been assigned the tags of "transportation" and "collision", but not for the neighboring city of Alexandria, VA where the data have been tagged with "transit" and "accident".

### 3.2 Schematic Challenges

The second challenge to civic data discovery is schematic. Schematic heterogeneity is understood to result from variations of conceptual schemata within or across disciplines; it can be overcome by schema integration [5].

Governmental organizations such as law enforcement agencies that report traffic accidents, including bicycle collisions, adhere to such integrated specifications, in this case the Model

<sup>4</sup> <http://lod-cloud.net/>

<sup>5</sup> <https://livingatlas.arcgis.com/en>

<sup>6</sup> <https://www2.usgs.gov/science/about/thesaurus-full.php>

Minimum Uniform Crash Criteria (MMUCC)<sup>7</sup> developed by the National Highway Traffic Safety Administration (NHTSA). This data model provides a reporting schema; local agencies can adapt it as needed, but it defines a minimum set of uniform fields that can be identified across municipal crash datasets. These criteria specify attribute names (i.e. “County Name”), definitions, and expected data types (i.e. “GLC Code”). Another well-adopted data model developed with interoperability in mind is the Local Government Information Model<sup>8</sup>. Similarly, it defines feature datasets (i.e. “Facilities Streets”), feature classes (i.e. “street lane width”), and attribute fields (i.e. “lane width, type: small integer”).

Where common data models are used, it is possible to easily reuse, and even combine, datasets. However, the majority of data discoverable through Esri’s ArcGIS Hub do not conform to any common data models. Attribute fields are defined ad-hoc and are also not indexed for search unless specified separately as tags.

### 3.3 Linked Data Approaches

The need for improved access to civic data parallels that for academic data. Just as research groups, or even academic domains, publish and reuse data according to different standards across various repositories, governmental agencies and municipalities also adhere to a variety of standards with varying levels of quality. The rise of Internet of Things (IoT) technology, which is enabling the evolution of “smart cities”, has also created new sets of challenges related to the volume, velocity, and variety of civic data streams. The challenges that have made heterogeneous civic data difficult to integrate and harmonize in the past have been successfully met by semantic annotation of data streams, which enables their alignment [3].

Rather than semantically annotating civic data after the fact, some governments have adopted linked open data principles as a standard for data sharing; “smart cities” such as London<sup>9</sup> and Dublin<sup>10</sup> have launched campaigns to expose operational city service data streams in an open, consumable format [7]. Esri Ireland for example now serves national geospatial information as linked data, consumable through an API [6]. In a linked data framework, it is not only easier for both humans and machines to consume civic data, but it is also easier to combine data from multiple sources, for example across levels of government.

One reason for this is that semantically annotated data can be dereferenced, resolving issues of uncertainty concerning attribute values or terms. For example, the United Nations Sustainable Development Goals ontology resolves terminological ambiguity while tracking progress toward shared goals on a multinational scale [11]. The outcomes of such successful linked data approaches motivated us to develop a similar method for semantically annotating civic data in order to improve user search. This method is the focus of the next section.

## 4 Methods

In order to improve the discoverability of civic data, we have developed and implemented a base vocabulary and a semantic annotation system. Semantic annotation augments official metadata with relevant tags supplied by a vocabulary, thus expanding the relevance recommendations of data. The method taken to develop and implement an automated semantic annotation system is summarized in the following steps:

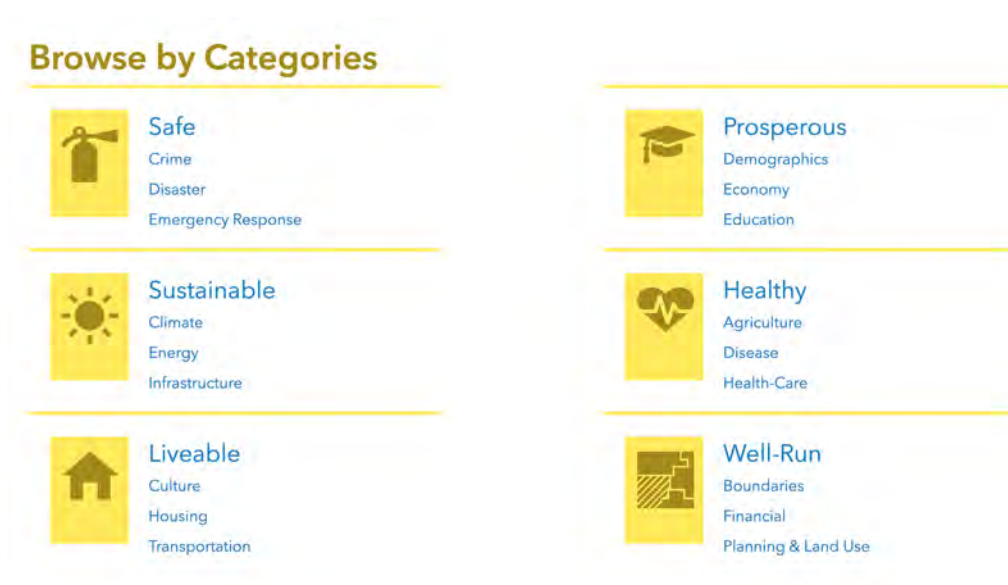
---

<sup>7</sup> <https://www.nhtsa.gov/mmucc>

<sup>8</sup> <http://solutions.arcgis.com/local-government/help/local-government-information-model/>

<sup>9</sup> <http://connected-data.london/>

<sup>10</sup> <http://smartdublin.ie/>



■ **Figure 1** ArcGIS Hub categories reflect existing themes assigned to datasets manually as tags.

1. Formalize base vocabulary for core geospatial themes
2. Extend vocabulary by reusing existing concept hierarchies
3. Augment existing metadata with extended tag hierarchies
4. Evaluate system performance for search

#### 4.1 Formalizing the Base Vocabulary

A key contribution of this work is the development and formalization of a compact base vocabulary that maps prototypical themes of government departments to aspects of users' lives. This vocabulary addresses two geospatial problems: 1) it makes data shared by governments that are co-located or adjacent discoverable; and 2) it makes descriptions of the phenomena that data are about semantically relevant to public users. The base vocabulary categories shown in Figure 1 were developed in collaboration with civic stakeholders, municipal staff, research organizations, and Esri's Local Government Team<sup>11</sup>. The vocabulary holistically organizes data and tools, allowing them to be referenced.

While these categories reflect typical organizational structures of civic government, they also capture core geospatial themes that communities want to track and measure. These categories are currently used as search facets for data in ArcGIS Hub. While they may structurally reflect issues that communities prioritize, they may not reflect the terms that community members may use when searching for this data. They also may not reflect the terms that a given organization uses to describe its data.

In order to formalize ArcGIS Hub Categories, we began by building a thesaurus of concepts modeled in Protégé<sup>12</sup>, an open source ontology editing software. We opted for a pragmatic adoption of the Simple Knowledge Organization System (SKOS) to model these concepts for a number of reasons: SKOS supports flexible modeling of hierarchical relationships; it is

<sup>11</sup><http://www.esri.com/software/arcgis/arcgis-for-local-government>

<sup>12</sup><https://protege.stanford.edu/>

used widely across numerous domains; and it is often used in term expansion activities<sup>13</sup>. For these reasons, we were able to reuse authoritative and dereferenceable concepts already published to the Semantic Web by organizations also using SKOS.

Some data available through ArcGIS Hub, such as layers exposed through Esri's Living Atlas of the World, have already been classified and tagged with ArcGIS Hub Categories. These include broader categories like "healthy" and narrower categories like "disease".

However, user-specified terms are not reflected in the ArcGIS Hub Categories. Analysis of the ArcGIS Hub query log revealed that users of Esri's ArcGIS Hub tend to search for data using terms that relate to their own colloquial conceptualizations of theme and geography. In a sample of 470,796 queries performed in 2015, only 12,257 (or 2.6 percent) used any form of the predefined categorical Hub keywords, (i.e. "healthy", "transportation", ...). This means that the majority of themes present in user searches likely take another form. This could mean that users are searching with synonyms of these keywords (i.e. "well-being"), or narrower concepts (i.e. "bicycle"), which would not yield results. Similarly, in the same sample of queries, only 64,353 (or 27.3 percent) use geographic references, like coordinates, addresses, place types, or zip codes in their searches. Similarly, geographic concepts that reflect place hierarchy (i.e. "Ronald Reagan National Airport is in Arlington County, VA") or proximity (i.e. "Reagan Airport is next to East Potomac Park") are not reflected in results.

## 4.2 Extending the Base Vocabulary

We imported existing concepts matching the Hub category tags from Library of Congress Subject Headings (LCSH)<sup>14</sup>, Princeton WordNet 3.1<sup>15</sup>, and the USGS Thesaurus<sup>16</sup>. Reusing these three vocabularies to describe civic data is novel, as they have been developed and traditionally used to describe library resources and scientific data. These vocabularies provide sufficient terminological coverage for extending the Hub categories shown in Figure 2.

LCSH are a controlled and well-defined set of terms used for resource classification. In addition to providing a stable identifier, LCSH concepts also adhere to a SKOS scheme and provide broader, narrower, and related concepts for each term. For example, "agriculture" in LCSH has useful variants "farming" and "husbandry", narrower terms like "agronomy", and related terms like "food supply" and "land use, rural". LCSH is designed to be used as a thesaurus; its subject headings provide bibliographic access to related subject matter.

Similarly, WordNet terms are also available in a SKOS scheme and are consumable as RDF, a linked data model. WordNet is a lexical database that combines the capabilities of a dictionary and a thesaurus for the English language. Concepts matching Esri Hub categories were retrieved from WordNet synsets, which are sets of synonyms with translations. For example, the synset for "agriculture" in WordNet includes "husbandry" and "farming" along with multilingual translations for each. Designed to support cognitive science applications, WordNet is suitable for information retrieval, text classification, and translation tasks [10].

A final source of Hub concept extension comes from the United States Geological Survey (USGS) Thesaurus, which is currently under development. As such, it provides identifiers without dereferencing; despite this, it is a rich source of authoritative scientific definitions and related terms in a SKOS scheme. For instance, it provides examples of the term agriculture used in the topics of "farming" and "horticulture". The USGS Thesaurus is designed to aid public interpretation of science web resources and topics.

<sup>13</sup><https://www.w3.org/TR/skos-ucr/>

<sup>14</sup><http://id.loc.gov>

<sup>15</sup><http://wordnet-rdf.princeton.edu>

<sup>16</sup><https://www2.usgs.gov/science/about/thesaurus-full.php>





■ **Figure 2** Extension of ArcGIS Hub terms to related categories in existing vocabularies.

Other sources were experimented with but ultimately were not implemented. Schema.org<sup>17</sup> was considered for thematic and geographic expansion, but was rejected as its top-level concepts are too broad, while narrowing too quickly. Geonames and DBpedia were also investigated, but have not yet been implemented; concepts from these sources may be included in the near future, as both are rich sources of colloquial place-types and themes found in users’ daily lives. It will be possible to extend the base vocabulary following the method developed in this work as other candidate vocabularies are considered.

To further expand ArcGIS Hub terms, we undertook additional mappings from existing categories to community standards, including INSPIRE<sup>18</sup>, FGDC<sup>19</sup>, and ISO 19115 data specifications. INSPIRE provides 34 spatial data themes, which specify common data models and code lists. INSPIRE themes aim to support the creation of a European Union spatial data infrastructure. These themes include “hydrography”, “transport networks”, and “protected sites”. Similarly, the National Geospatial Data Asset (NGDA) provides a set of 16 themes with appointed lead agencies and the aim of supporting data interoperability. These themes include “climate and weather”, “land use-land cover”, and “soils”. Finally, ISO 19115 provides a set of 19 themes, including terms like “biota”, “health”, and “oceans”. Each of these community standards function as a controlled vocabulary for describing spatial data resources in their respective metadata contexts; their terms overlap to varying extents.

Pragmatically, we were interested in areas of term overlap, as mapping these standardized community terms to the expanded set of ArcGIS Hub terms establishes semantic links between thematically related resources. Various agencies conform to these standards when describing their data. Federal agencies, such as the U.S. Geological Survey, use NGDA themes to describe resources shared through ArcGIS Hub. The FGDC for example maintains a keyword thesaurus with these terms and points to it as a best-practices resource for publishing

<sup>17</sup> <http://schema.org/docs/schemas.html>

<sup>18</sup> <https://inspire.ec.europa.eu/data-specifications/2892>

<sup>19</sup> <https://www.fgdc.gov/what-we-do/manage-federal-geospatial-resources/a-16-portfolio-management/themes>

The image shows two screenshots from a web interface. The left screenshot is titled 'SPARQL query' and shows a query template with various fields like 'EXAMPLE QUERIES', 'PREFIXES', 'SPARQL ENDPOINT', and 'CONTENT TYPE'. The right screenshot is titled 'Term Expander' and shows the expanded terms for 'agriculture' in JSON format, including synonyms and translations.

**SPARQL query**

To try out some SPARQL queries against the selected dataset, enter your query here.

EXAMPLE QUERIES

Selection of triples   Selection of classes

PREFIXES

rdf   rdfs   owl   xsd   **o**

SPARQL ENDPOINT:    CONTENT TYPE (SELECT):

```

2 PREFIX hub: <http://www.esri-hub.com/vocab/>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
6 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
7 PREFIX ontology: <http://wordnet-rdf.princeton.edu/ontology#>
8 PREFIX loc: <http://ld.loc.gov/authorities/subjects/>
9
10 SELECT ?label ?synonym
11 WHERE {
12   ?term skos:prefLabel ?label .
13   ?term skos:altLabel ?synonym .
14   FILTER contains( lowercase(?label), "agriculture" )
15 }

```

**Term Expander**

Expands a term using SPARQL query against Fuseki

Term:   

```

{
  "synonyms": [
    "Farming",
    "Husbandry",
    "agribusiness",
    "agricultura",
    "agroindustria",
    "factory farm"
  ],
  "translation": [
    "الزراعة الحضرية",
    "الزراعة",
    "الزراعة",
    "agricultura",
    "landbrug",
    "laborantza",
    "nekezaritza",
    "الزراعة",
    "maanviljelyst",
    "maanviljelyst",
    "maatloos",
    "agriculture",
    "cucuk tanam",
    "pengebunan",
    "penternakan",
  ]
}

```

■ **Figure 3** SPARQL query template (left) and expanded terms (right) for term “agriculture”.

datasets to open data clearinghouses; it states that “the more robust your theme keyword list, the more likely it can be located by others (and yourself)”. While this is true in principle, describing data with controlled keywords alone will not make data readily discoverable for public consumers of data who often search for data using colloquial terminology.

In order to augment official metadata, the controlled vocabularies for INSPIRE, NGDA, and ISO 19115 were incorporated into the expanded ArcGIS Hub terms. We designated mappings between related terms from each controlled vocabulary in Protégé using the SKOS predicate *related*. Thus, a term like “transportation” has: *related* terms from INSPIRE (“Transport networks”), NGDA (“Transportation”), ISO 19115 (“Transportation”); *broader* and *narrower* terms from LCSH and USGS Thesaurus (“public transit”); and *synonyms* and *translations* from WordNet (“ES - transporte”). Each of these tags becomes a triple statement pointing to externally defined resources.

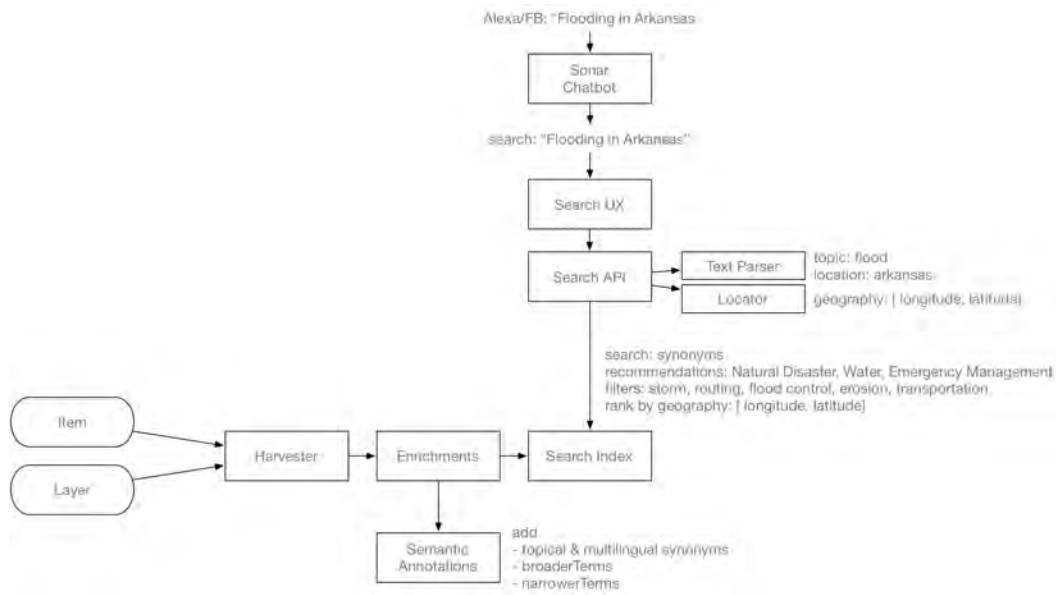
### 4.3 Augmenting Existing Metadata

We exported the base vocabulary from Protégé as triple statements in Terse RDF Triple Language (Turtle)<sup>20</sup> syntax and imported them into a Fuseki<sup>21</sup> triplestore, set up as a public endpoint. The vocabulary is stored as a graph that can be queried using SPARQL syntax, which allows for queries across multiple endpoints. Figure 3 shows an example of a query template in Fuseki returning query results in JSON to be integrated as auxiliary metadata.

ArcGIS Hub includes a search index of aggregated dataset records from all data providers. When organizations like governments indicate their data is public, ArcGIS Hub compiles multiple metadata sources into a custom search index to support multiple content search and discovery services.

<sup>20</sup> <https://www.w3.org/TeamSubmission/turtle/>

<sup>21</sup> <https://jena.apache.org/documentation/>



■ **Figure 4** Semantic annotations added to metadata, supporting search through query expansion.

The search index process, shown in Figure 4, includes three phases: harvesting, validation, and enrichment. During harvesting of a dataset, Hub collects metadata from the ArcGIS Online item information, associated formal metadata, the feature service and feature layer definition, and data attribute aggregate statistics. Validation includes heuristics to measure metadata completeness, support for secure connections with HTTPS, and query responsiveness, which determines if the data are actually accessible. During the enrichment phase, a dataset is decomposed into relevant keywords which are then sent to the semantic query service to retrieve new semantic tags that are then attached to the dataset metadata.

For example, Flood Zone data from Evansville, Indiana are tagged “Evansville, Vanderburgh County, Flood Zones, IN, environmental”. Using each of the terms from each of the tags results in a superset of synonyms, translations, broader terms, and narrower terms, shown in Figure 5. These terms are each added to the dataset record in the search index using an internal semantic annotation service. The semantic annotation service is an internal API that hosts the base vocabulary as a queryable API using the Apache Jena Fuseki server. This server supports defined requests to build a set of tags that expand the dataset metadata for broad, narrow, translated, and similar terms.

At query time, these additional terms can be used to match user queries such as “human health”, or “impact assessment” that may not have another similar word match in the dataset metadata collection but will now have results based on matching these new, additional semantic tags. The semantic tags also include translations such as “air pasang” (Indonesian) or “nousuvesi” (Finnish). Beyond similar terms, there are broader terms such as “Natural disasters”, and “Water” and narrower terms such as “Flood damage prevention” and “Forest influences” that can be used to recommend new search terms to the user for refining their search results.

## 4.4 Evaluating system performance for search

In practice, search for data is now semantically aided; related content, such as synonymous terms, can be retrieved when inferred as thematically related. For example, a search for traffic accidents can now return other content related to a broader concept of ‘transportation’ as pedestrian fatalities. While only a small fraction of data (about 8 percent) in ArcGIS Hub initially included formal metadata, semantic annotations added related metadata in the form of related terms, supporting data discoverability and integration.

In order to evaluate the contributions of our approach, we consider that semantically enabled search wasn’t previously possible: this informs our baseline criteria. Search efficacy is measured accordingly using several methods: conversion rates through usage analytics tracking, usability testing, and relevance judgment evaluation.

Usage analytics tracking measures all user interactions with the ArcGIS Hub web application. This includes search inputs, filter interactions and result selection. We define several conversion funnels corresponding to expected user outcomes, which include downloading the data, creating an information product such as a web map or a Story Map, or bookmarking a view of the data for later use. These conversions indicate that a good search result was returned. We can then compare conversion results with and without semantic annotation.

Usability testing includes defining a workflow that human test subjects perform while being monitored by researchers. Listening to stream of consciousness verbal evaluations and observing interface interactions denotes perceptions of different search modalities and outcomes. This testing may be performed in-house or in collaboration with stakeholders.

Lastly, relevance judgment evaluation asks a similar set of users to evaluate the quality of search results as: perfect, relevant, partially relevant, or irrelevant. The scores for each result are tallied and compared with the optimal result and rank ordering to define the quality of the search relevance, due to semantic annotations or without semantic annotations.

The results of these evaluation measures are forthcoming at this time of writing.

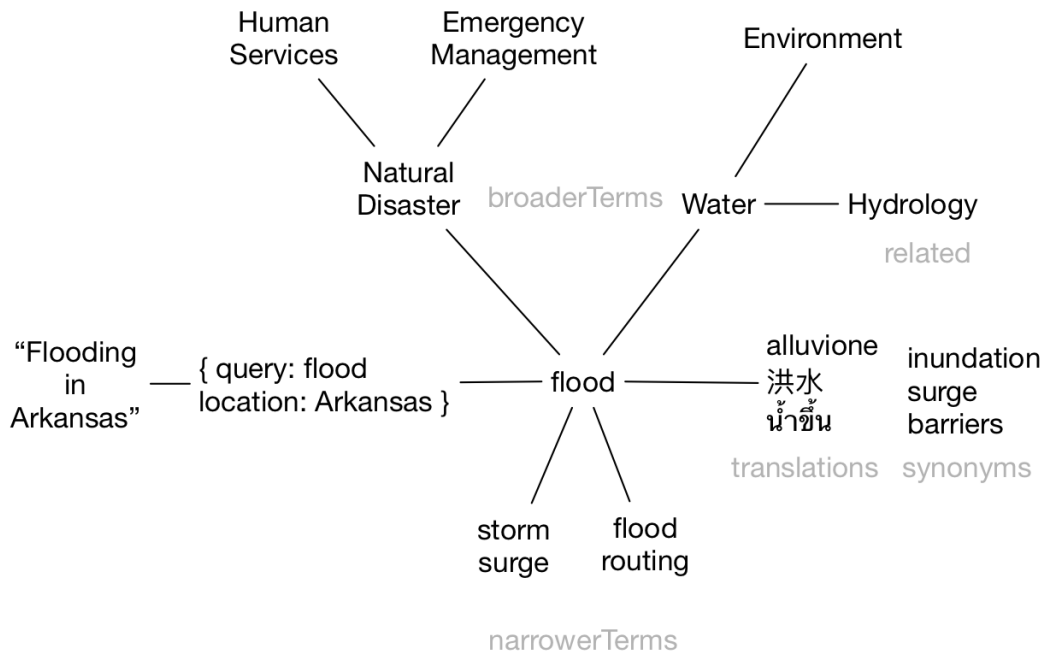
## 5 Results

Governments, academic institutions and other organizations publish open data to encourage the creative reuse of information for new purposes. ArcGIS Hub allows these organizations to create websites that enable search and discovery of their authoritative data, as well as recommend data shared through other groups. The Bureau of Transportation’s Geospatial Statistics site is shown as one such example in Figure 6. Visitors can perform simple searches through their web browser or mobile device, or request information through new digital media chatbots on Facebook and Amazon Alexa.

Extending dataset metadata with semantic annotations expands the discoverability of information through colloquial and multilingual search associations. Figure 4 illustrated how search queries use the semantic search index to parse and retrieve relevant datasets.

To use the semantic search API, Hub implements a REST HTTP API for structured queries from web browsers, mobile apps, and custom embeds; it uses a JSON-Schema self-documenting hypermedia API and includes search index attribute filters and facets. An API search query is first split into relevant parameters for keywords, time, location, and provider. The keywords are compared with the semantic annotation tags for similarity matches; the time, location and provider are used as filters. The result includes a relevance-ranked list of datasets as well as aggregate facets of topics, data types, and providers for further filtering.

The semantic annotations augment the search relevance matches by comparing search keywords with terms that may not have existed in the original metadata document, but describe the dataset with alternative labels that match these queries. Figure 5 shows an



■ **Figure 5** Search queries are parsed and compared with semantic annotations to expand matches and provide additional facets.

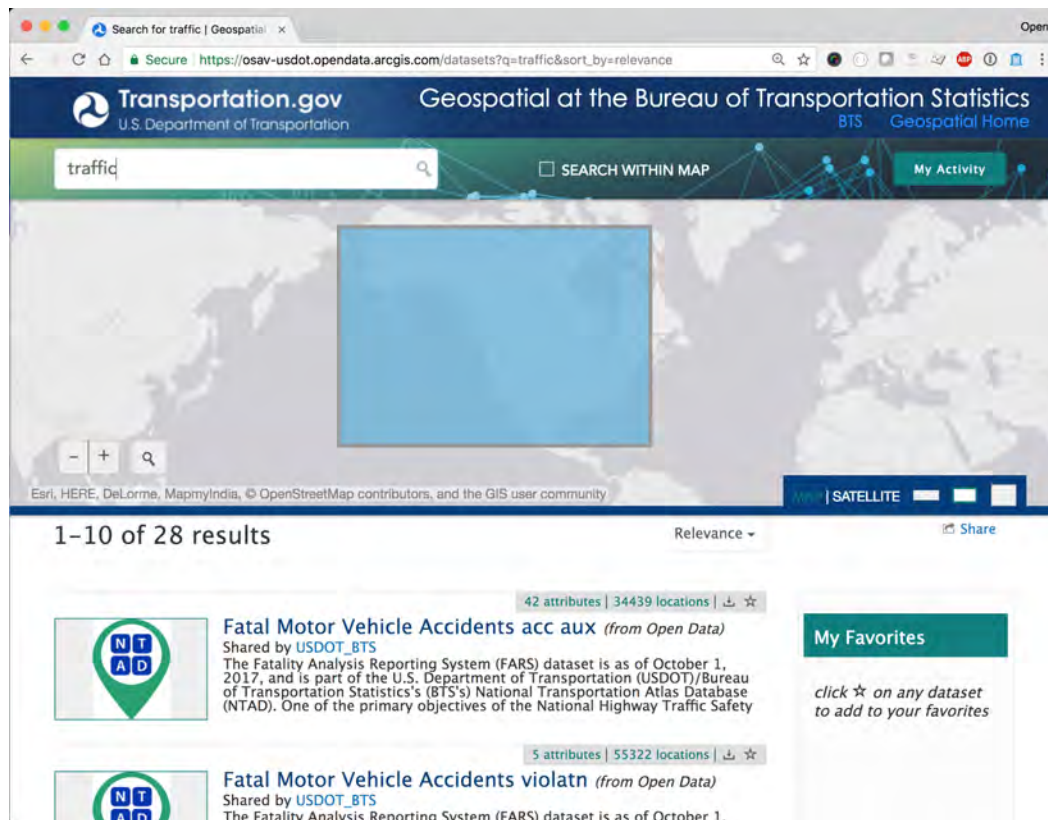
example of expanded semantic tags that are compared for relevance ranking, including multilingual terms, as well as the broader and narrower aggregate terms that can be used in search interface facets.

### 5.1 Building Data Networks

Semantic annotation supports additional use cases beyond metadata querying. ArcGIS Hub includes a global catalog of data from governments of various administrative levels: local council and departmental, metropolitan, provincial, regional, national, and multinational organizations. Each government follows a varying set of metadata and keyword standards that may not overlap with other governments, even if the organizations are geographically adjacent or coincident. This can make integration of data across municipal boundaries problematic, resulting in lost productivity or detriments to operations and safety.

Semantic annotations support data integration by organizing datasets into common thematic groupings, which increase the discovery and utilization of similar datasets across municipal data providers. By way of example, consider several civic datasets provided by neighboring municipalities such as road networks, public schools, moving violations (e.g. vehicle speeding citations), and reported crashes between vehicles, bicycles or people. Additionally, there are regional and national datasets provided by agencies that also include transit networks (bus stops and train stations): FARS (Fatality Analysis Reporting System).

In order to track progress toward thematic community initiatives, such as “Vision Zero”, discovery of relevant data must be possible across all levels of government. Vision Zero is a strategy to eliminate all traffic fatalities and severe injuries, while increasing safe, healthy, equitable mobility for all. Potential Federal data sources for tracking a “Vision Zero” Initiative are shown in Figure 6. However, without semantic annotation, there is uncertainty as to



■ **Figure 6** U.S. Department of Transportation traffic related datasets sorted by relevance.

whether a search for traffic data will return relevant results across other Hub sites at a state, county, or municipal level.

ArcGIS Hub builds the search index that includes each of the four example local municipal datasets from each municipality. This includes the original metadata and the additional semantic annotations on the datasets that associate them with related thematic groupings. Searching just the category term has mixed, or missing, results from some provider catalogs. Figure 7 compares search results across the GIS catalogs of the District of Columbia, State of Maryland, and County of Arlington, exposed through ArcGIS Hub.

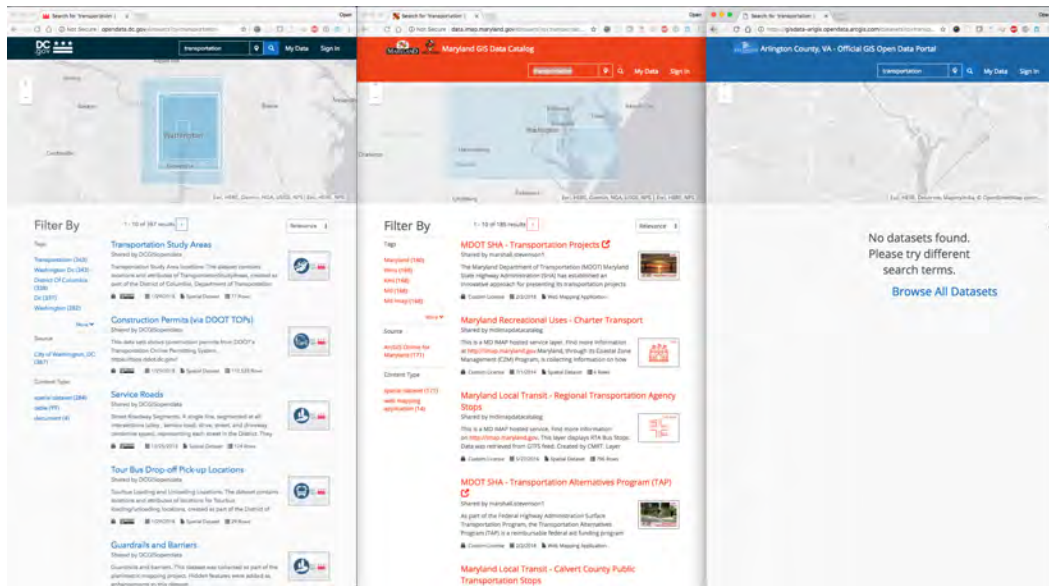
By comparison, when colloquial terms are used, there are similar results from all local providers. Figure 8 compares search results across the same GIS catalogs for related terms.

## 6 Discussion and Outlook

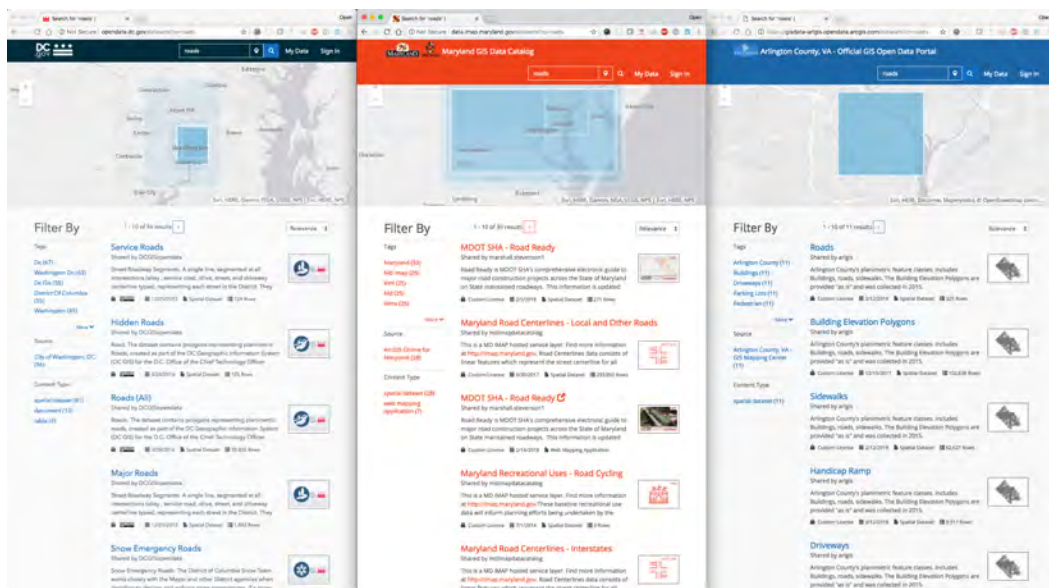
The work presented in this paper improves data discovery through the application of semantic annotations to civic data, which facilitate transparency and coordination of work; semantic search enables the exploration and discovery of relationships among organizations' data that were previously unknown.

Several areas of research are continuing from this work. We plan to expand and refine the base vocabulary to better support bi-directional term expansion. This will allow users to discover new datasets by improving traversal of the base vocabulary's relations, like broad and narrow terms, for both thematic and geographic concepts. We anticipate that alignment with new ontologies, such as the U.N.'s Sustainable Development Goals Ontology,

9:14 Improving Civic Data Discovery



■ Figure 7 Comparing searches for “transportation” before adding semantic annotations.



■ Figure 8 Comparing searches for related term “roads” after adding semantic annotations.

and application of our methods in related domains, such as academic libraries, will continue to improve data discovery across organizational repositories.

On a larger scale, the lessons learned from our research can be applied to new domains and extended along the following dimensions.

The Sustainable Development Goals (SDGs) are the results of an ambitious global initiative to improve the health and well-being of people and communities. They consist of 17 goals, 169 targets and 232 data indicators that will measure and monitor progress towards the SDG. These targets and indicators include a semantic graph that relate to socioeconomic terms, municipal planning, and other related governance sectors. Work is ongoing with

several national mapping agencies and the United Nations to integrate their semantic graphs with the base vocabulary presented in this paper.

We are also applying the methods developed in this paper to data discovery in the context of digital research libraries. While libraries have long been the traditional brokers of knowledge, today's queries are largely mediated by commercial digital search engines [12]. Yet, libraries are taking on new roles, facilitating discovery, and often co-production, of knowledge [8]. Semantically annotated data can be more easily discovered and retrieved via queries that traverse knowledge graphs, regardless of the endpoints where they are hosted. Academic libraries are poised to serve as a semantically-neutral meeting ground where domain data can be aggregated and made spatially and thematically discoverable, similar to ArcGIS Hub.

---

## References

- 1 Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*, 2010. doi:10.1038/npre.2010.4626.1.
- 2 Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.
- 3 Stefan Bischof, Athanasios Karapantelakis, Cosmin-Septimiu Nechifor, Amit P Sheth, Alessandra Mileo, and Payam Barnaghi. Semantic modelling of smart city data. In *Report of the W3C Workshop on the Web of Things 2014*, 2014. URL: <https://www.w3.org/2014/02/wot/papers/karapantelakis.pdf>.
- 4 Wade Bishop and Tony H Grubestic. Geographic information, maps, and gis. In *Geographic Information*, pages 11–25. Springer, 2016.
- 5 Yaser Bishr. Overcoming the semantic and other barriers to gis interoperability. *International journal of geographical information science*, 12(4):299–314, 1998.
- 6 Christophe Debruyne, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, and Declan O'Sullivan. Serving ireland's geospatial information as linked data. In *International Semantic Web Conference (Posters & Demos)*, 2016.
- 7 Rob Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.
- 8 Sara Lafia, Jon Jablonski, Werner Kuhn, Savannah Cooley, and F Antonio Medrano. Spatial discovery and the research library. *Transactions in GIS*, 20(3):399–412, 2016.
- 9 Matthew S Mayernik. Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4):973–993, 2016.
- 10 George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- 11 Barry Smith and Mark Jensen. The unep ontologies and the obo foundry. In *ICBO/BioCreative*, 2016.
- 12 Elaine Svenonius. *The intellectual foundation of information organization*. MIT press, 2000.
- 13 Open Research Data Taskforce. Research data infrastructures in the uk : Landscape report. Technical report, Universities UK, 2017.
- 14 Andrew Turner. Desire paths to open data. <http://highearthorbit.com/articles/desire-paths-to-open-data>, 2014.
- 15 Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014.