Transactions in GIS

Research Article Transactions in GIS, 2015, 19(3): 398–416

Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online

Yingjie Hu,* Krzysztof Janowicz,* Sathya Prasad† and Song Gao*

*STKO Lab, Department of Geography, University of California, Santa Barbara, California, USA

†Applications Prototype Lab, ESRI Inc., Redlands, California, USA

Abstract

Geoportals provide integrated access to geospatial resources, and enable both authorities and the general public to contribute and share data and services. An essential goal of geoportals is to facilitate the discovery of the available resources. Such a process relies heavily on the quality of metadata. While multiple metadata standards have been established, data contributers may adopt different standards when sharing their data via the same geoportal. This is especially the case for user-generated content where various terms and topics can be introduced to describe similar datasets. While this heterogeneity provides a wealth of perspectives, it also complicates resource discovery. With the fast development of the Semantic Web technologies, there is a rise of Linked-Data-driven portals. Although these novel portals open up new ways to organize metadata and retrieve resources, they lack effective semantic search methods. This article addresses the two challenges discussed above, namely the topic heterogeneity brought by multiple metadata standards and the lack of established semantic search in Linked-Data-driven geoportals. To harmonize the metadata topics, we employ a natural language processing method, namely Labeled Latent Dirichlet Allocation (LLDA), and train it using standardized metadata from Data.gov. With respect to semantic search, we construct thematic and geographic matching features from the textual metadata descriptions, and train a regression model via a human participants experiment. We evaluate our methods by examining their performances in addressing the two issues. Finally, we implement a semantics-enabled and Linked-Data-driven prototypical geoportal using a sample dataset from Esri's ArcGIS Online.

1 Introduction

Geoportals are online gateways that provide integrated access to geospatial resources, such as maps, services, and a variety of geo-data (Maguire and Longley 2005). As unified platforms, geoportals enable both authorities and public users to publish and share geospatial resources. Geoportals also form a key component of Spatial Data Infrastructures (SDIs) which facilitate access to geographic information and reduce data duplication among different government departments (Masser 1999).

An essential goal of geoportals is to facilitate the discovery of the available resources. Most geoportals maintain a catalog service which indexes the metadata of the resources in the portal (Lutz and Klien 2006). As depicted in Figure 1, the process of resource discovery typically follows three steps: (1) resource providers publish their metadata to the geoportal; (2) consumers

Address for correspondence: Yingjie Hu, STKO Lab, Department of Geography, University of California Santa Barbara, Santa Barbara, CA 93106, USA. E-mail: yjhu.geo@gmail.com

Acknowledgements: This work is a collaborative effort from the UCSB STKO Lab and the Esri Applications Prototype Lab. The authors would like to thank Jack Dangermond, Hugh Keegan, and Dawn Wright, as well as the three anonymous reviewers for their constructive comments and feedback.

 doi: 10.1111/tgis.12151

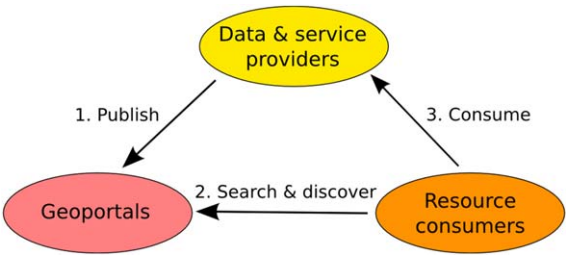

Figure 1 Typical *publish-find-bind* pattern in geoportals

search and potentially discover these resources through geoportals; and (3) consumers access the resources directly from the providers (Ostensen and Smits 2002). Two major factors can impact the performance of resource discovery: the quality of the metadata and the capability of the portal's search functionality.

Metadata play an important role in the resource discovery process. Accurate and complete metadata ease the discovery of relevant resources, whereas it is difficult to find data when metadata are incomplete or even missing. To ensure metadata quality, standards, such as the Content Standard for Digital Geospatial Metadata (CSDGM) (Lee and Chan 2000) from the Federal Geographic Data Committee (FGDC) as well as the ISO 19115 standard (Nogueras-Iso et al. 2004) have been developed. However, given the variety of data contributors, their motivations, cultural backgrounds, time constraints, and so forth, today's metadata vary dramatically in terms of quality and completeness. With the emergence of community-based geoportals, the general public also began to contribute geospatial resources (De Longueville 2010). Unlike specialists who have often received professional training, general users may not follow any specific metadata standard but only annotate their data with informal descriptions and tags. Semantically harmonizing these metadata facilitates the indexing of resources and also enables a more accurate categorization suitable for faceted search (Nowak and Craglia 2006; Craglia and Annoni 2007).

The search functionality is another important factor that influences the performance of resource discovery in geoportals. Traditionally, keyword-based search has often been employed to find resources based on the user's query. While achieving fair performance, keyword-based search often suffers from low recall, i.e. it fails to discover the resources which are described in semantically-relevant but syntactically-different terms (e.g. *road* versus *street*) (de Andrade et al. 2014). One such example can be found on *Data.gov* (Figure 2). As can be seen, searching the keyword *earthquake* returns more results than searching the keyword *natural disasters*, although earthquakes are generally considered as a type of *natural disasters*. To overcome this limitation, researchers proposed semantic search which retrieves data based on not merely keyword matching but semantic similarity (Janowicz et al. 2011). Ontologies, as semantic mediators among different concepts, have often been used to implement semantic search (Jones et al. 2004; Lutz and Klien 2006; Wiegand and Garcia 2007; Fox et al. 2009; Li et al. 2014; Janowicz et al. 2011).

With the advances in Semantic Web technologies and the rise of Linked Data, it is worth revisiting the metadata and search aspects that have hampered the effective use of geoportals for many years. The term *Linked Data* has two meanings which have been used interchangeably. On one side, it refers to a set of principles, recommended by W3C, to publish and share data on the Semantic Web. On the other side, it has also been used to refer to the data which have been published following these W3C principles. A *Linked-Data-driven geoportal* is a new type of geoportal which manages metadata not in traditional databases but by following Linked Data principles. The Resource Description Framework (RDF) is the standard method for such Linked-

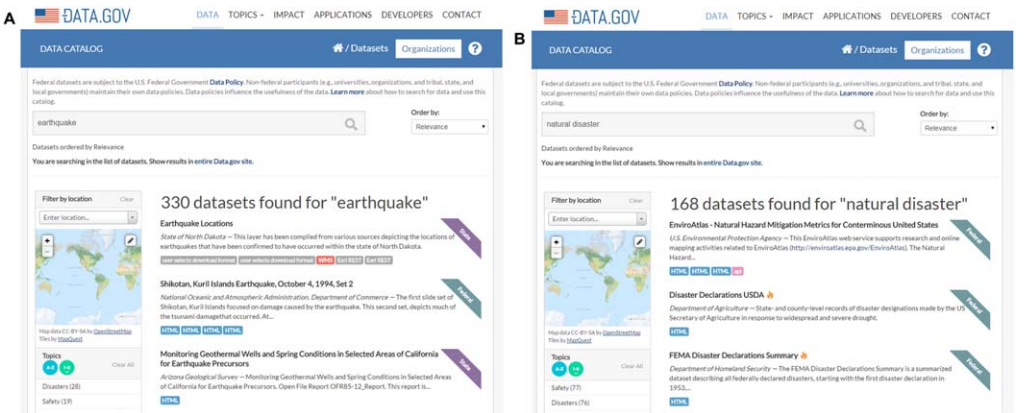

Figure 2 Limitation of keyword-based search using an example from Data.gov: (a) Search for "earthquake"; and (b) Search for "natural disaster"

Data-driven geoportals, and the metadata, represented using RDF, are stored as nodes and links which constitute a globally interconnected data graph. Such graph-based metadata organization possesses novel features, such as better accommodation of domain-specific attributes, deductive reasoning, global identifiers, and intuitive data browsing by following typed links (Athanasis et al. 2009; Zhang et al. 2010; Zhao et al. 2014). However, such new geoportals still lack effective semantic search methods which can function on the metadata represented as RDF.

This article addresses the two challenges discussed above, namely heterogeneous metadata caused by multiple standards as well as the lack of established semantic search in Linked-Data-driven geoportals. For the first challenge, we focus on the heterogeneous metadata topics which traditionally have to be harmonized manually. We employ a natural language processing method, namely Labeled Latent Dirichlet Allocation (LLDA), and train it using the standardized metadata from Data.gov. We then apply the learned LLDA model to the metadata from a major geoportal, i.e. Esri's ArcGIS Online, and automatically harmonize the metadata into the ISO 19115 standard based on unstructured textual descriptions. With respect to the semantic search, we construct thematic and geographic matching features from the textual descriptions, and train a regression model to rank the search results. Finally, we implement a semantics-enabled and Linked-Data-driven geoportal using a sample of ArcGIS Online metadata. The contributions of our research are as follows:

- We present a LLDA-based workflow to harmonize and enrich metadata. We benchmark our approach against a naive Bayesian classifier baseline.
- We develop a regression model to rank the search results based on the semantically-expanded input query. Such a model can be integrated into a Linked-Data-driven geoportal using a SPARQL query to support the semantic search functionality.
- We implement a prototypical geoportal using the proposed methods and a sample of metadata from Esri's ArcGIS Online.

The remainder of this article is structured as follows. Section 2 provides a review of the related work on geoportals, metadata standards, and Linked-Data-driven portals. Section 3 describes the methods for metadata harmonization as well as semantic search. Section 4 applies the presented methods to the experimental data retrieved from Data.gov and ArcGIS Online, and

evaluates their performance. Section 5 presents the prototype which has been implemented using data from ArcGIS Online. Finally, Section 6 summarizes this work and discusses future directions.

2 Related Work

The value of geoportals has been recognized by many governments and more than 100 geoportals have been established during the past decade (Rose 2004). On the global level, the GEOSS (Global Earth Observation System of Systems) portal (http://www.geoportal.org/web/guest/geo_home_stp) is an international effort from more than 70 countries to facilitate the integration and sharing of diverse environmental datasets and decision support tools (Onsrud et al. 2010). On the continental level, INSPIRE (Infrastructure for Spatial Information in the European Community) (http://inspire-geoportal.ec.europa.eu/) is the largest spatial data infrastructure in Europe that manages data from the European Union (Masser 2007). On the country level, there is Data.gov (https://www.data.gov/) which is the nationwide portal for the US that incorporates data from the previous Geospatial One-Stop (GOS) (Tait 2005) as well as other federal agencies (e.g. NOAA and USGS), and local governments. On the state level, a variety of portals have been developed, such as the Californian geoportal (http://portal.gis.ca.gov/geoportal/catalog/main/home) and other state-level portals, which share and maintain local datasets. In addition, there are a variety of domain-specific portals, such as FEMA's geoportal for disaster response (Walker and Maidment 2006), the South Carolina Community Assessment Network for public health (Tang and Selwood 2005), and the Greenhouse Gas Data Portal for climate change (Lin et al. 2013).

The fast growth of SDI also witnessed the evolution of metadata standards. In the US, FGDC is a major organization that coordinates the implementation of metadata standards. For many years, The Content Standard for Digital Geospatial Metadata (CSDGM) has been used as the standard to organize metadata (Maguire and Longley 2005). With the emergence of ISO 19115, FGDC has began to encourage the use of this international standard. On its official website (http://www.fgdc.gov), FGDC states that *"federal agencies are encouraged to transition to ISO metadata as their agencies are able to do so"* and that *"It's recognized that the transition to ISO metadata will be occurring over the next few years."* (retrieved in January 2015). To facilitate metadata transition, conversion tools, such as the XML transformation from NOAA's National Coastal Data Development Center Initiative, have been developed (Nogueras-Iso et al. 2004). These conversion tools establish mappings between the CSDGM metadata elements and the corresponding elements in ISO 19115. However, there is a mandatory element in ISO 19115, namely the topic category, which cannot be directly converted from the existing metadata (FGDC 2006). Although it is possible to ask data providers to check their data and manually add suitable topic categories, such a process is labor-intensive and time-consuming. Alternatively, semantic alignment approaches (Euzenat 2004; Giunchiglia et al. 2007) could be employed to align the keywords or themes from the original metadata to the ISO 19115 topics. For example, Fugazza and Vaccari (2011) developed SKOSMatcher, a tool that enables domain experts to align terms in independent structured vocabularies. In the EuroGEOSS project, a semantic broker (http://www.eurogeoss-broker.eu/) was created to interlink thesauri in different languages so that multilingual queries can be supported. While good at mapping controlled vocabularies, semantic alignment approaches have limited performance in handling informal or free metadata keywords that do not conform to existing formal vocabularies. This research adopts a machine learning approach to automatically enrich metadata with ISO 19115 topics. Such an approach makes use of the

commonly available textual descriptions (e.g. titles and snippets) in metadata, and therefore is not restricted to controlled vocabularies. Currently, our research focuses on metadata in English.

Linked-Data-driven geoportals are a new type of portal constructed using the Semantic Web technologies and Linked Data principles. One important advantage of a Linked-Data-driven geoportal is its capability to index not only pre-defined attributes but also domain-specific information. For example, in traditional database management systems, a data provider who publishes a hurricane dataset may not be able to get the dates or strengths of the hurricanes indexed, since this information is domain-specific and does not belong to an existing metadata standard. However, Linked-Data-driven geoportals are based on the flexible use of ontologies, and therefore can address this issue by including corresponding domain-specific ontologies. Athanasis et al. (2009) developed an early example of such a Linked-Data-driven geoportal to accommodate the different attributes from multiple domains. Keßler et al. (2012) developed a Linked Data portal for the GIScience community to organize geospatial data related to researchers. In addition, a Linked-Data-driven geoportal allows users to intuitively browse data by following links between resources. Such a feature can be helpful when users are uncertain about what queries to input and would like to explore the data by "following their noses". Given metadata organized as RDF graphs, users can navigate from one resource (e.g. a map) to a related resource (e.g. a data layer used in this map). However, semantic search functionality for the hosted RDF metadata has not been realized to date.

3 Methods

3.1 Harmonizing Metadata Topics

One geoportal may store metadata from a wide variety of data contributers. Consequently, different metadata standards may be used (if used at all), and the metadata quality may vary. In such cases, it is important to harmonize the heterogeneous metadata. As ISO 19115 is an international standard recommended by FGDC, this work will try to harmonize heterogeneous metadata into this standard. Specifically, we will focus on assigning topic categories which are a mandatory element in ISO 19115.

We develop a machine learning based workflow for this topic categorization task. Such a workflow considers this task as a multi-label classification problem, i.e. one geospatial resource can be assigned multiple topics. Such a consideration is derived from our empirical experience with geospatial datasets which usually have more than one topic. For example, a dataset about *Pollution from vehicles* can have both *environment* and *transportation* as its thematic topics. Similarly, NASA's ASTER Global Digital Elevation Map (GDEM) can be categorized under the topics of both *elevation* and *imageryBaseMapsEarthCover* using ISO 19115. We formalize the topic categorization problem as below:

Problem: *Given a metadata record m, its textual description d, and a set of topics T: $\{t_1, t_2, ..., t_n\}$, assign a subset of topics T' to m so that each topic t_i in T' has a relevance score s_i based on d that is larger than a threshold τ.*

For each metadata record m, its textual description d is the major data used by our workflow to decide which topics to assign. Such a design can enhance the generalizability of the developed workflow, since textual descriptions can be commonly found in metadata in the form of titles and short descriptions (often referred to as *snippets*). The raw textual descriptions will go through a sequence of natural language processing steps, including removing

punctuation and stop words, checking case consistency, and stemming words. The processed results are used as input features for the learning model.

The relevance score s_i between the description d and a topic t_i needs to be calculated using a selected model. The model needs to be first tuned using training data, and can then be applied to the unseen data. Depending on the specific problems, different training datasets should be selected. For example, if the goal is to harmonize metadata topics following the standard of ISO 19115, then metadata which are constructed in ISO 19115 need to be used as the training data. On the other hand, if the goal is to harmonize the metadata into another standard, a different set of training data in that standard should be used. As we are handling a text-based classification problem, a common approach is to use naive Bayesian models. However, we utilize LLDA in this work, and will explain the reasons in the following paragraphs. Figure 3 provides an overview of the workflow using an example of the ISO 19115 standard.

Naive Bayesian models are a common approach for text-based classification. They consider each textual description as a *bag of words*, and estimate the probability that a textual description belongs to a category by multiplying the word-based posterior probabilities:

$$P(t_i|d) \propto \prod_{j=1}^{N} P(w_j|t_i) \times P(t_i) \tag{1}$$

where $P(t_i|d)$ is the posterior probability that a given textual description d belongs to the topic t_i; $P(t_i)$ is the prior probability of topic t_i in the training dataset; and $P(w_j|t_i)$ is the posterior probability that given the topic t_i, word w_j will appear in the textual description.

While naive Bayesian models have been successfully used in many existing text-based classification problems (Rennie 2001; Sebastiani 2002), they have several limitations when applied to harmonizing metadata topics in this work. First, naive Bayesian models are not suitable for multi-label classification. While they can be used to generate multiple topics, naive Bayesian models assume that each textual description is associated with only a single topic when calculating the posterior probability (see Equation 1). This assumption undermines the capability of naive Bayesian models in modeling the textual descriptions of geospatial resources, which are often associated with multiple topics. Secondly, naive Bayesian models use the empirical prior probabilities estimated from the training data as the prior probability for unseen data. This is often not the case for the metadata harmonization problem, since the data to be harmonized may come from geoportals that have different percentages of topics from those of the training dataset. Finally, naive Bayesian models consider each single word as an input feature, and can lead to overfitting for long textual descriptions. In Equation (1), N represents the number of words in a textual description, and the value of N can vary significantly for different descriptions, thereby affecting the classification result.

Labeled Latent Dirichlet Allocation (LLDA) is a probabilistic graphical method which models the process for generating a textual document involving one or multiple topics (Ramage et al. 2009). LLDA is developed based on the traditional unsupervised Latent Dirichlet Allocation (LDA) model (Blei et al. 2003), but introduces a supervised component. Like LDA, LLDA models a document (a textual description in our case) as a mixture of topics, and each of these topics has a learned probabilistic distribution over a vocabulary from the training dataset. Unlike LDA, LLDA constrains the learned topics to be those observed in the training dataset, instead of using a group of words to represent each topic as LDA does. This design overcomes the limitation of LDA in which some learned topics are difficult to interpret. Figure 4 shows a graphic representation of this probabilistic model.

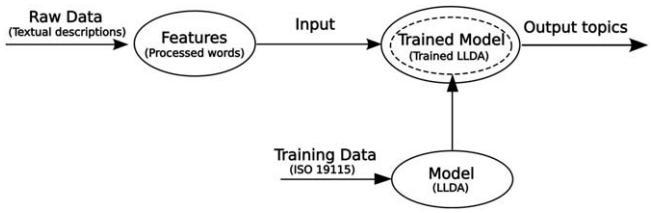

Figure 3 An overview of the metadata harmonization workflow using an example of the ISO 19115 standard

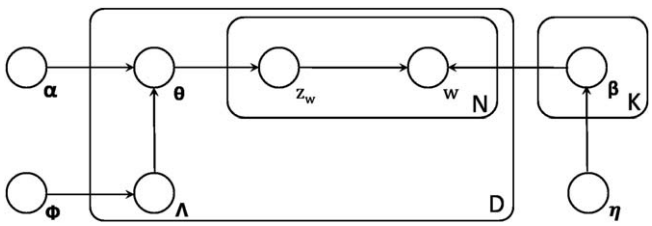

Figure 4 Graphic representation of the LLDA model: Both the topic set $\mathbf{\Lambda}$ and the prior probabilities derived from $\boldsymbol{\alpha}$ influence the topic mixture $\boldsymbol{\theta}$ (Ramage et al. 2009)

In Figure 4, D represents a set of documents, and we use d to represent a single document in D. K is a set of topics in the training data, and in the example of the ISO 19115 standard, K contains 19 topics, such as *farming*, *biota*, *elevation*, and so forth. N represents a set of words in document d. $\mathbf{\Phi}$ is a vector of parameters for the Bernoulli distribution which generates $\mathbf{\Lambda}$. $\mathbf{\Lambda}$ is a topic presence indicator $\{l_1, l_2, ..., l_K\}$, where $l_k \in \{0, 1\}$ (1 indicates the presence of a topic and 0 otherwise). $\boldsymbol{\alpha}$ is a vector of parameters for a Dirichlet distribution which contributes to the prior probabilities of topic mixture $\boldsymbol{\theta}$. $\boldsymbol{\eta}$ is a vector of parameters for another Dirichlet distribution which generates $\boldsymbol{\beta}$ for each topic k. $\boldsymbol{\beta}$ represents the prior probabilities for each word in the vocabulary that belongs to the topic k. z_w is a generated topic from $\boldsymbol{\theta}$, and it contributes to the generation of the word w. The parameters $\boldsymbol{\alpha}$, $\mathbf{\Phi}$, and $\boldsymbol{\eta}$ are learned from the training dataset, and the trained LLDA model can then be applied to the unseen data. For more details on LLDA, please refer to Ramage et al. (2009).

For the problem of harmonizing metadata topics, LLDA overcomes the limitations of naive Bayesian models. First, LLDA models a textual description as a mixture of topics, and multiple topics have been taken into account simultaneously. Such a modeling process fits the problem of multi-label classification. Meanwhile, LLDA possesses a higher generalizability on unseen data. This is because the prior probabilities of topics are modeled as random variables following a Dirichlet distribution, instead of as single values purely based on the frequency of training data (as naive Bayesian models do). Finally, LLDA makes use of dimensionality reduction to combine word-based features into a fixed set of new features, and avoids the issue of overfitting for long textual descriptions. In addition, the dimensionality reduction based on linear combination of words can also capture some degree of synonymy and polysemy (Blei et al. 2003).

3.2 Semantic Search for RDF Metadata

This section presents the semantic search method we have developed for Linked-Data-driven geoportals in which metadata are represented as RDF. Specifically, we adopt a free-text-style

search which allows users to input any textual query instead of restricting the input terms to certain vocabularies. Also, we enable users to perform geospatial queries using place names. Such a design can be helpful when the users know the name of the target place but are unfamiliar with its location or boundary.

3.2.1 Metadata enrichment

As our goal is to build a free-text-style search, the textual descriptions in metadata can be useful sources against which the text-based query can be semantically compared. The descriptions in the original metadata often contain information that cannot be directly used for search, such as punctuations and stop words. Thus, our first step is to extract meaningful information from the descriptions, and insert the extracted result as RDF triples back into the metadata. By manually examining a sample of metadata from Data.gov, we found that thematic concepts (e.g. *population* and *earthquake*) and geographic names (e.g. *California* and *Utah*) in textual descriptions can often represent the major content of the corresponding geospatial resources. Based on this observation, we take a named entity recognition (NER) approach to enrich metadata.

NER requires a background knowledge base against which the concepts and entities can be extracted. We employ DBpedia as our knowledge base, which is a Semantic Web version of Wikipedia and which contains the entities and concepts in Wikipedia articles (Auer et al. 2007; Lehmann et al. 2014). A two-stage procedure has been used for the NER process: spotting and disambiguating. In the spotting stage, terms which can be used to represent concepts and geographic names are identified from the textual descriptions. These terms are obtained from three sources of DBpedia: article titles, redirect pages, and disambiguation pages. In the disambiguating stage, the identified terms are disambiguated based on the surrounding context words using a vector space model and cosine similarity measurement. More details about this two-stage procedure have been discussed in previous studies (Mendes et al. 2011; Hu et al. 2014).

In many existing geoportals, such as Data.gov, INSPIRE, and Esri's ArcGIS Online, there are often two types of textual descriptions, namely titles and snippets. Titles usually provide more concise and meaningful information than the snippets. Therefore, we differentiate the concepts extracted from titles and snippets using four vectors:

$$[\mathbf{T}_t, \mathbf{T}_g, \mathbf{S}_t, \mathbf{S}_g] \tag{2}$$

where $\mathbf{T}_t$ is a vector that contains the thematic concepts extracted from the title; $\mathbf{T}_g$ represents the geographic names from the title; and $\mathbf{S}_t$ and $\mathbf{S}_g$ are the other two vectors representing the concepts and geographic names extracted from the snippet. The four vectors are then inserted back into the RDF metadata using the SPARQL statement as below.

```
INSERT data {
 :exampleResource :hasTitleThematicTerm :titleThematicTerm .
 :exampleResource :hasTitleGeoTerm :titleGeoTerm .
 :exampleResource :hasSnippetThematicTerm :snippetThematicTerm .
 :exampleResource :hasSnippetGeoTerm :snippetGeoTerm . }
```

Listing 1 SPARQL statement for inserting thematic concepts and geographic names into the metadata

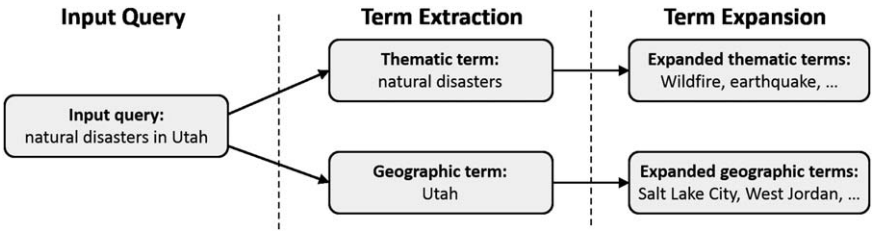

Figure 5 The query expansion process using an example query of *natural disasters in Utah*

3.2.2 Query expansion

With the enriched metadata, our second step for semantic search is to expand the user's input query to help discover relevant resources. Existing studies suggest that thematic concepts and geographic names should be expanded differently (Jones et al. 2001; Fu et al. 2005; Li et al. 2011). For thematic concepts, the expansion should focus on including syntactically-different but semantically-similar words. For geographic names, the expansion needs to consider the place relevance and hierarchies.

Ontologies, such as SWEET (Semantic Web for Earth and Environmental Terminology), have often been used to expand thematic concepts in geoportals (Lutz and Klien 2006; Li et al 2011; de Andrade et al. 2014). However, such an approach also restricts the terms that can be used in the input queries, and therefore does not meet our goal of designing a free-text-style search. The latent semantic analysis (LSA) can discover hidden relations between terms using a smoothing window and a mathematical method called singular value decomposition (SVD) (Landauer et al.1998). By analyzing a large amount of text, LSA can find semantically related words for almost any given term. In a recent work, Han et al. (2013) trained a LSA model based on the content from 100 million web pages. Instead of repeating existing work, we embed their trained LSA model into our system, and use this model to expand thematic concepts. For geographic names, existing gazetteers usually contain useful information about the relevance and hierarchy of places. For fast implementation, we make use of the Geonames gazetteer service (http://api.geonames.org), which contains a rich volume of data about geographic places. Figure 5 illustrates the process of query expansion using an example query of *natural disasters in Utah*. It is worth noting that thematic concepts and geographic names are first extracted from the input query, and are then expanded. The concept extraction is conducted using the same two-stage procedure employed in the metadata enrichment.

3.2.3 Result ranking

With enriched metadata and the expanded query, our third step is to quantify the relevance between geospatial resources and the input query, and to rank the result based on the calculated relevance values. As represented in Equation (2), the textual descriptions in each metadata record have been represented as four vectors representing the thematic and geographic concepts extracted from the title and the snippet, respectively. Meanwhile, since the input query has been expanded to include similar terms, we also differentiate *exact match* from *similar match*. *Exact match* means the vectors contain a term which is exactly

the same as the original input query, whereas *similar match* indicates that the vectors contain a term which is not in the original query but is added in the query expansion process. As each of the four vectors can contain both *exact match* and *similar match*, we generate eight matching scores as below.

In addition to the eight matching scores, we also introduce an interaction variable, called *Thematic-Geo Interaction* (TGI), which has been defined as follows:

$$TGI=(TTE+TTS+STE+STS)\times(TGE+TGS+SGE+SGS) \tag{3}$$

As can be seen, TGI is the multiplication between the sum of thematic matching scores and the sum of geographic matching scores. The rationale for introducing this interaction variable is that a good search result on geospatial resources often needs to have both thematic and geographic matches. Consider a user input query *Drugs and Crime in California*. A map that has very high thematic matching score but is about *Drugs and Crime in Spain* (thus 0 geographic matching score) may not necessarily be of interest to the user. On the contrary, a map that has low scores in both thematic and geographic matching, e.g. *Robberies in Los Angeles*, may be considered as a good match by the user. Thus, neither the thematic nor the geographic scores have a fixed influence on the relevance value; instead the influence of one type of scores depends on the existence of the other type, and the TGI variable, as denoted by its name, captures this interaction. In fact, we will test the value of the interaction variable in the evaluation experiment.

Table 1 Matching features constructed for ranking

Title Thematic Exact match (TTE)	Title Thematic Similar match (TTS)
Title Geographic Exact match (TGE)	Title Geographic Similar match (TGS)
Snippet Thematic Exact match (STE)	Snippet Thematic Similar match (STS)
Snippet Geographic Exact match (SGE)	Snippet Geographic Similar match (SGS)

Based on the 9 scores, a simple regression model has been developed to quantify the relevance between an input query and a candidate resource.

$$\begin{aligned} R(q,m)=&\lambda_1 TTE+\lambda_2 TTS+\lambda_3 TGE+\lambda_4 TGS \\ &+\lambda_5 STE+\lambda_6 STS+\lambda_7 SGE+\lambda_8 SGS+\lambda_9 TGI \end{aligned} \tag{4}$$

where $R(q,m)$ represents the relevance between the query q and the metadata record m. $\lambda_1, \lambda_2, ..., \lambda_9$ are weights for the matching scores, and we will show how to estimate these parameters in the evaluation experiments. Please note that we have designed this regression model as not having an intercept, since the relevance score should be 0 when no match exists. Thus, we force the regression line to pass through the origin, and make the intercept on the y axis as 0.

The presented regression model can be integrated into a Linked-Data-driven geoportal using one SPARQL query. Such a SPARQL query is formalized as below:


```
SELECT ?item (COUNT(?titleThematicExact) AS ?TTE
(COUNT(?titleThematicSimilar) AS ?TTS)
(COUNT(?titleGeoExact) as ?TGE)
(COUNT(?titleGeoSimilar) as ?TGS)
(COUNT(?snipThematicExact) as ?STE)
(COUNT(?snipThematicSimilar) as ?STS)
(COUNT(?snipGeoExact) as ?SGE)
(COUNT(?snipGeoSimilar) as ?SGS)
(((?TTE+?TTS+?STE+?STS)*(?TGE+?TGS+?SGE+?SGS)) as ?TGI)
(( λ1*?TTE + λ2*?TTS + λ3*?TGE + λ4*?TGS + λ5*?STE + λ6*?STS +
λ7*?SGE + λ8*?SGS + λ9*?TGI) as ?ranking)
WHERE {
  OPTIONAL {
   ?item :hasTitleThematicTerm ?titleThematicExact .
   FILTER ( ?titleThematicKey = :exactThematicTerm ) }
  OPTIONAL {
   ?item :hasTitleThematicTerm ?titleThematicSimilar .
   FILTER ( ?titleThematicSimilar = :expandedThematicTerm ) }
  OPTIONAL {
   ?item :hasTitleGeoTerm ?titleGeoExact .
   FILTER ( ?titleGeoExact = :exactGeoTerm ) }
  OPTIONAL {
   ?item :hasTitleGeoTerm ?titleGeoSimilar .
   FILTER ( ?titleGeoSimilar = :expandedGeoTerm ) }
  OPTIONAL {
   ?item :hasSnippetThematicTerm ?snipThematicExact .
   FILTER ( ?snipThematicExact = :exactThematicTerm ) }
OPTIONAL {
 ?item :hasSnippetThematicTerm ?snipThematicSimilar .
 FILTER ( ?snipThematicSimilar = :expandedThematicTerm ) }
OPTIONAL {
 ?item :hasSnippetGeoTerm ?snipGeoExact .
 FILTER ( ?snipGeoExact = :exactGeoTerm ) }
OPTIONAL {
 ?item :hasSnippetGeoTerm ?snipGeoSimilar .
 FILTER ( ?snipGeoSimilar = :expandedGeoTerm ) }
 } GROUP BY ?item ORDER BY Desc(?ranking) LIMIT 1000
```

Listing 2 SPARQL query for calculating the relevance scores and ranking the result based on the regression model

4 Evaluation Experiments

4.1 Experimental Data

The experimental data are retrieved from two geoportals: Data.gov and ArcGIS Online. Data.gov maintains a catalog service which provides metadata constructed using the ISO 19115 standard. In total, we retrieved 26,917 metadata records from Data.gov. ArcGIS Online is a community-based geoportal which contains resources not only from federal agencies but also contributed by the general public. The metadata from ArcGIS Online have varied qualities:

they do not conform to existing ISO or FGDC standards, most records are not annotated using controlled vocabularies, and some attribute values are even missing. In total, 10,201 metadata records have been retrieved from ArcGIS Online.

The overall experimental procedure is as follows: (1) we use the metadata from Data.gov to train our LLDA model, and evaluate its performance by comparing it with a naive Bayesian model; (2) we apply the trained LLDA to the unstandardized ArcGIS Online data to automatically assign ISO 19115 topics, and we implement the semantic search using the proposed methods; and (3) we tune the weights of the presented regression model using a human participants experiment, and evaluate the quality of the search results.

4.2 *Evaluation for Metadata Topic Harmonization*

The data from Data.gov contain the ISO 19115 topics assigned by data providers, and therefore have been considered as the ground truth in this experiment. In order to compare the performances of the LLDA and the naive Bayesian model, we train both models using the same Data.gov dataset. The 10-fold cross validation has been employed for this evaluation. Such a method divides the data into 10 folds, and iteratively trains and tests the model; in each iteration, it uses nine folds of data for training and the remaining one fold for testing; finally, the performances in all the iterations are summarized to provide an overall score of the model (Flach 2012). To quantify the performance, we employ two metrics that have been commonly used in information retrieval, namely precision and recall, which have been defined in Equations (5) and (6). We then calculate the precision and recall for both the LLDA and the naive Bayesian model, and plot out the result in Figure 6.

$$precision = \frac{retrieved\ relevant}{all\ retrieved} \tag{5}$$

$$recall = \frac{retrieved\ relevant}{all\ relevant} \tag{6}$$

Precision and recall are a trade off: a high precision often comes with low recall and vice versa. However, as can be seen in Figure 6, at any given point of recall, LLDA shows a higher precision than the naive Bayesian model, demonstrating a generally better performance. When the trained LLDA model is put into use, choosing a high precision with low recall indicates that most topics generated by the model are correct, but there may be quite a number of metadata entries that the model "feels uncertain about" and therefore cannot assign a topic. On the contrary, choosing a high recall with low precision indicates that the model "is bold enough" to assign topics to most metadata, but cannot guarantee the correctness of these topics. Depending on the requirements of specific applications, different thresholds τ can be used to achieve the desired precision and recall. When applying the trained LLDA to the unstandardized ArcGIS Online data, we choose a threshold 0.37 which achieves 0.80 for precision and 0.69 for recall.

To understand the possible reasons for the misclassification, we manually examine the misclassified records and compare them with the ground truth topics assigned by data providers. It has been found that the model makes a number of mistakes when the metadata records only have very short textual descriptions. Such mistakes are understandable since the model has only limited information to make a decision. In addition, there are some cases in which proper topics generated by the LLDA model are considered as errors since these topics are not assigned by data providers. For example, we find a metadata record on *Greenhouse gas data* which was assigned with topics of *environment* and *climatologyMeteorologyAtmosphere* by the LLDA model. While

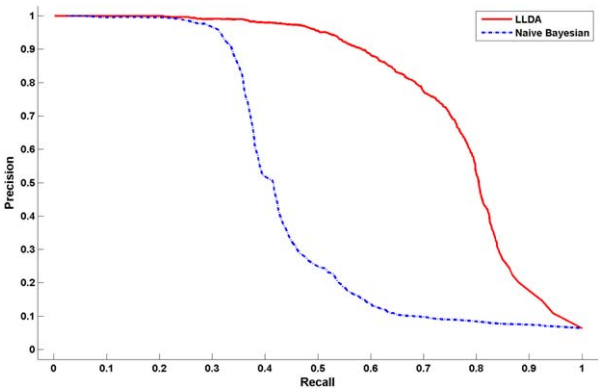

Figure 6 The precision-recall plot of the LLDA model compared with the naive Bayesian model; the red solid curve represents the result from the LLDA; the blue dotted curve represents the result from the naive Bayesian model

both topics make sense, the data provider has only assigned *environment* to this metadata record, and therefore the topic *climatologyMeteorologyAtmosphere* has been considered as a mistake.

4.3 Evaluation for the Semantic Search

To evaluate our proposed method for semantic search, we conduct a human participants experiment in which seven individuals were invited to evaluate the search results based on 10 input queries. For each query, we provide a search phrase (e.g. "california population density") and 10 candidate results with varied relevance to the input query. Each participant was asked to judge the degree of relevance from zero (not matching at all) to five (perfect matching). To help ensure reproducibility, the experimental queries and their candidate maps can be downloaded at http://stko-exp.geog.ucsb.edu/survey/Questions.zip. The experiment results and calculated scores can be downloaded at http://stko-exp.geog.ucsb.edu/survey/surveyResult.zip. In total, we have collected 700 data records from this experiment, and these human judgments have been considered as our ground truth.

The 10-fold cross validation has been adopted again to train and test our regression model for the semantic search. We first average the judgments from the seven different individuals for each query and each candidate map, and obtain an aggregated dataset with 100 data records (10 queries and each query has 10 candidate maps). We then iteratively train the regression model, and use the trained model to estimate the relevance values for the testing data records. Finally, the estimated relevance values are compared with the averaged judgments from the participants. Pearson's correlation coefficient (Equation 7) has been used to quantify the comparison result:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{7}$$

We plot out the estimated relevance scores and the human judgments in Figure 7. To test the value of the interaction variable, two regression models (with interaction and without interaction) have been fitted using the training data, and their estimated scores on the testing data have been shown in Figures 7a and b, respectively. The solid diagonals in the plotes represent a perfect consistency between the estimated scores and human judgments, and the dotted lines are

fitted based on the actual data points which have human judgments as the x coordinates and estimated scores as the y coordinates. In a perfect case, all of the data points should fall onto the diagonals (since the estimated scores should equal the human judgment scores), and the dotted line should overlap with the solid line. The more the dotted line deviates from the solid diagonal, the worse the model's performance is. As can be seen, including the interaction variable brings a higher correlation coefficient 0.7226 ($P < 0.001$), compared with the regression model without the interaction variable. Meanwhile, the dotted line in Figure 7(b) is closer to the solid diagonal than the dotted line in Figure 7(a), indicating that the model with the interaction variable achieves a generally better consistency between the estimated scores and human judgments.

5 Prototype

5.1 Implementation

As a proof-of-concept, we have implemented a semantics-enabled and Linked-Data-driven geoportal using the experimental metadata retrieved from ArcGIS Online. This prototype can be accessed at: http://stko-exp.geog.ucsb.edu/linkedportal/. The ArcGIS Online metadata have been converted into RDF, and are hosted in a SPARQL endpoint. For more details about the RDF conversion, entity naming, and data publishing, readers can refer to our previous work (Hu et al. 2015). The LLDA model trained based on Data.gov data has been used to automatically assign ISO 19115 topics to the unstructured ArcGIS Online metadata. The regression model has been tuned using the data from our human participants experiment, and has been embedded into the geoportal using the SPARQL query proposed in Listing 2. To increase the efficiency of the SPARQL query, a block of union statements has been dynamically inserted into the SPARQL query to preselect the resources which have at least one matched concept before the execution of the OPTIONAL statements. An example of the union block is shown in Listing 3. Based on this implementation, a query submitted to our geoportal prototype can typically be responded within five seconds.

```
SELECT ?item (COUNT(?titleThematicExact) AS ?TTE
  ...
WHERE {
 /* begin the UNION block */
 { ?item :hasTitleThematicTerm :titleThematicExact }
 UNION { ?item :hasTitleGeoTerm :titleGeoExact }
 UNION { ?item :hasSnippetThematicTerm :snipThematicExact }
 ...
 UNION{ ?item :hasSnippetGeoTerm :snipGeoSimilar }
 /* finish the UNION block */

 OPTIONAL {
 ?item :hasTitleThematicTerm ?titleThematicExact .
 FILTER ( ?titleThematicKey = :exactThematicTerm ) }
 OPTIONAL {
 ?item :hasTitleThematicTerm ?titleThematicSimilar .
 FILTER ( ?titleThematicSimilar = :expandedThematicTerm ) }
 ... } GROUP BY ?item ORDER BY Desc(?ranking) LIMIT 1000
```

Listing 3 A block of union statements inserted into the SPARQL query to preselect the geospatial resources that have at least one match

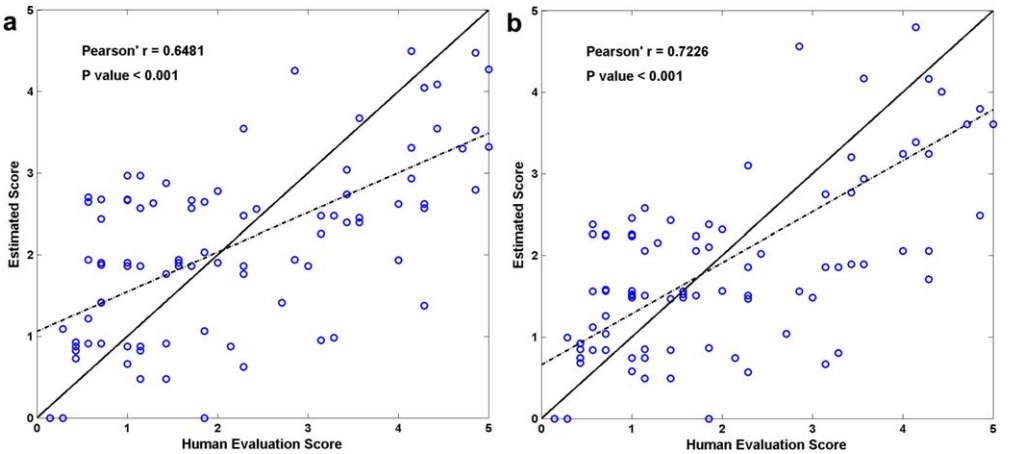

Figure 7 Comparing estimated relevance scores with human judgments: (a) Without the interaction variable; and (b) With the interaction variable

The metadata harmonization provides standard ISO 19115 topics based on which the semantic search results can be grouped into different topic categories. This thematic categorization allows facet search, i.e. the users can not only see all the returned results but can also browse them by topic categories. To implement facet search, we design a set of tabs on the left side of the user interface, and dynamically summarize the number of the returned items for each topic. In addition, we also allow users to browse the returned items by thematic matching, geographic matching, or having both matchings. In the following subsections, we employ two scenarios to show the major functions of the implemented geoportal.

5.2 Free-Text-Style Semantic Search

The implemented geoportal allows users to input queries freely instead of restricting the search terms to pre-defined vocabularies. When a query is submitted to the geoportal, it will be processed by a Web service which analyzes the query and extracts the thematic and geographic terms. Such terms are expanded and enriched with related thematic concepts and geographic names, and are then compared with the candidate map records using the SPARQL query. In this scenario, we repeat a previous query executed on Data.gov: searching "natural disaster" and "earthquake". Two screenshots of the search results have been shown in Figure 8. It can be seen that a search of "natural disaster" returns maps which are about wildfire, hurricane, earthquake, and other natural disasters and which do not necessarily have the keyword "natural disaster".

5.3 Link-tracing Data Browsing

Since the metadata in the geoportal are hosted in RDF, users can browse data through their links. By clicking at one of the search results (which are represented as cards with thumbnails), users can see detailed information about this geospatial resource in the displayed metadata table. Meanwhile, the highlighted links in the metadata table can lead users to other related resources. For example, users can click the link of the map owner to find out who has created this map and other public information about the map creator. Similarly, users can also click at other highlighted links, such as the layers used in the map, to explore the data in the geoportal.

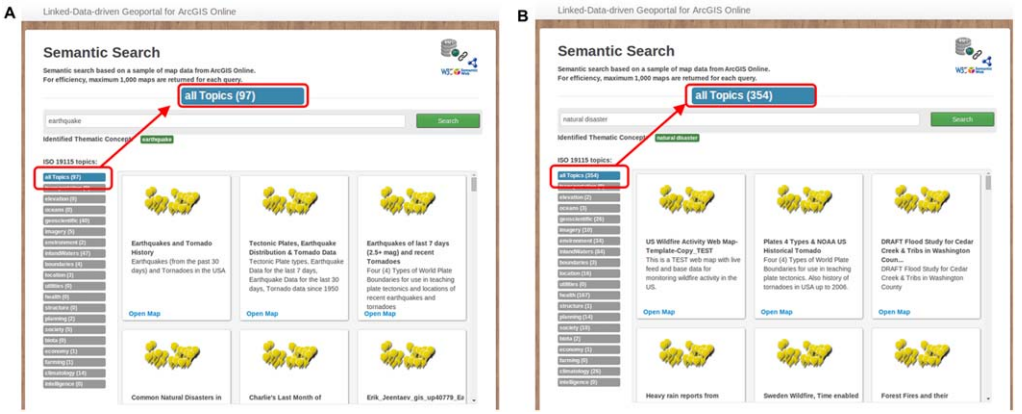

Figure 8 Free-text-style semantic search on the implemented geoportal: (a) Search for "earthquake"; and (b) Search for "natural disaster"

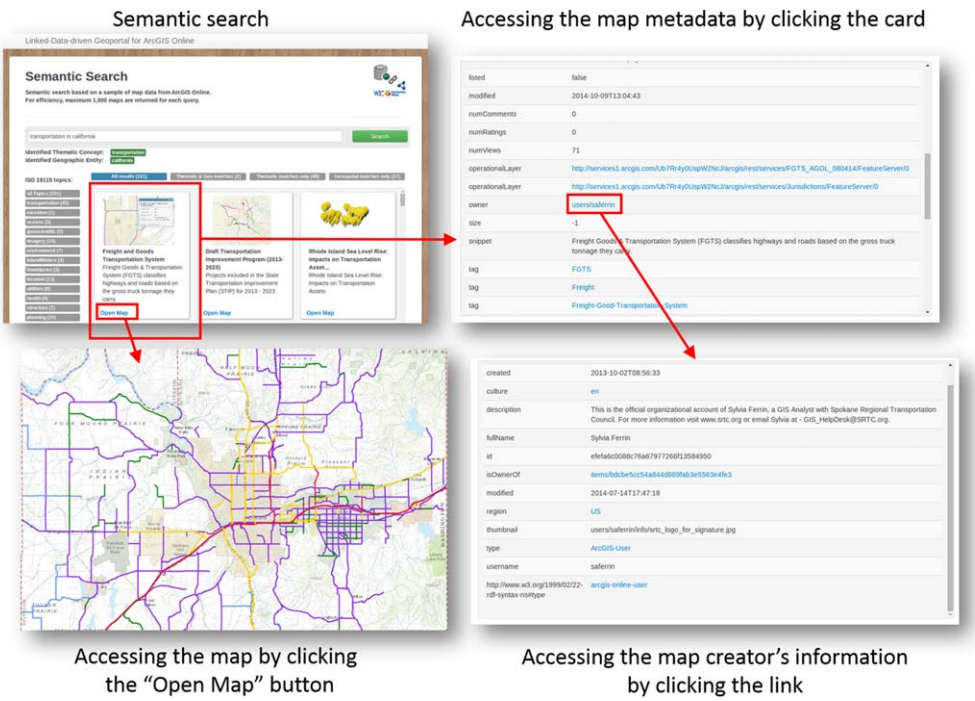

Figure 9 Browsing resources by tracing data links

Besides, by clicking the "Open Map" button, one can directly access the geospatial resource published on ArcGIS Online. Figure 9 demonstrates this link-tracing browsing process.

6 Conclusions and Future Work

A major function of geoportals is to facilitate the discovery of geospatial resources. Today's Big Data age brings a huge amount of data whose volume and variety are increasing at an

unprecedented speed. Under this context, there is a growing demand for methods which can facilitate the resource discovery in geoportals. Our article is an effort in this direction. Specifically, we focus on two issues which can influence the resource discovery result, namely metadata topic heterogeneity brought by multiple standards and the lack of established semantic search in Linked-Data-driven geoportals. For the metadata heterogeneity, we train a LLDA model to automatically assign topics to metadata records, and semantically harmonize them. Our evaluation result shows that LLDA has an overall better performance compared with that of a naive Bayesian model. Linked-Data-driven geoportals use RDF to organize and interlink metadata into graphs, and possess several merits including accommodating domain-specific metadata and enabling link-tracing data browsing. For the lack of semantic search on RDF data, we develop methods for metadata enrichment and query expansion, and construct thematic and geographic matching features to quantify the relevance between an input query and the candidate resources. We then train a regression model using a human participants experiment, and rank the search results based on the constructed matching scores. Our method achieves fair performance in the evaluation experiment, and can be integrated into an existing Linked-Data-driven geoportal using one SPARQL query.

This research, however, still has limitations that could be improved in future work. First, the weights in the regression model are derived from a small-scale human participants test, and therefore could be biased. A larger sample of participants could be recruited to achieve a more accurate ranking result. In addition, the semantic search function still takes several seconds to return in our current implementation. This response time could significantly increase when a large amount of metadata have been inserted into the system. Thus, we still need to examine the scalability of the system and improve its efficiency in later work.

References

Athanasis N, Kalabokidis K, Vaitis M, and Soulakellis N 2009 Towards a semantics-based approach in the development of geographic portals. *Computers and Geosciences* 35: 301–08

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, and Ives Z 2007 DBpedia: A nucleus for a web of open data. In *Proceedings of the Sixth International Semantic Web Conference*, Busan, South Korea: 722–35

Blei D M, Ng A Y, and Jordan M I 2003 Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022

Craglia M and Annoni A 2007 INSPIRE: An innovative approach to the development of spatial data infrastructures in Europe. In Onsrud H J (ed) *Research and Theory in Advancing Spatial Data Infrastructure Concepts*. Orono, ME, Global Spatial Data Infrastructure Association: 93–105

de Andrade F G, de Souza Baptista C, and Davis C A 2014 Improving geographic information retrieval in spatial data infrastructures. *GeoInformatica* 18: 793–818

De Longueville B 2010 Community-based geoportals: The next generation? *Computers, Environment and Urban Systems* 34: 299–308

Euzenat J 2004 An API for ontology alignment. In McIlraith S A, Plexousakis D, and van Harmelen F (eds) *The Semantic Web: ISWC 2004*. Berlin, Springer Lecture Notes in Computer Science Vol. 3298: 698–712

FGDC 2006 ISO 19115 Topic Categories from ISO/DIS 19115; Metadata Quick Guide. WWW document, http://www.fgdc.gov/metadata/documents/MetadataQuickGuide.pdf

Flach P 2012 *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. New York, Cambridge University Press

Fox P, McGuinness D L, Cinquini L, West P, Garcia J, Benedict J L, and Middleton D 2009 Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. *Computers and Geosciences* 35: 724–38

Fu G, Jones C B, and Abdelmoty A I 2005 Ontology-based spatial query expansion in information retrieval. In Tari Z (ed) *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*. Berlin, Springer Lectures Notes in Computer Science Vol. 3761: 1466–82

Fugazza C and Vaccari L 2011 Coupling human- and machine-driven mapping of SKOS thesauri. *International Journal of Metadata, Semantics and Ontologies* 6(3): 155–65

Giunchiglia F, Yatskevich M, and Shvaiko P 2007 Semantic matching: Algorithms and implementation. *Journal on Data Semantics* 9: 1–38

Han L, Kashyap A, Finin T, Mayfield J, and Weese J 2013 UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, Atlanta, Georgia: 4–52

Hu, Y., Janowicz, K., Prasad, S., 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *Proceedings of the Eighth ACM SIGSPATIAL Workshop on Geographic Information Retrieval*, Dallas, Texas: 1–8.

Hu Y, Janowicz K, Prasad S, and Gao S 2015 Enabling semantic search and knowledge discovery for ArcGIS Online: A linked-data-driven approach. In *Proceedings of the Eighteenth AGILE International Conference on Geographic Information Science*, Lisbon, Portugal

Janowicz K, Raubal M, Kuhn W 2011 The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science* 2: 29–57

Jones C B, Abdelmoty A I, Finch D, Fu G, and Vaid S 2004 The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Proceedings of the Third International Conference on Geographic Information Science (GIScience 2004)*, College Park, Maryland: 125–39

Jones C B, Alani H, and Tudhope D 2001 Geographical information retrieval with ontologies of place. In Montello D R (ed) *Spatial information theory: Foundations of Geographic Information Science International Conference, COSIT 2001 Morro Bay, CA, USA, September 19–23, 2001 Proceedings*. Berlin, Springer Lecture Notes in Computer Science Vol. 2205: 322–35

Keßler C, Janowicz K, and Kauppinen T 2012 spatial@linkedscience: Exploring the research field of GIScience with linked data. In Xiao N, Kwan M-P, Goodchild M F, and Shekhar S (eds) *Geographic Information Science: Seventh International Conference, GIScience 2012, Columbus, OH, USA, September 18–21, 2012, Proceedings*. Berlin, Springer Lecture Notes in Computer Science Vol. 7478: 102–15

Landauer T K, Foltz P W, and Laham D 1998 An introduction to latent semantic analysis. *Discourse Processes* 25: 259–84

Lee Y and Chan H 2000 Spatial metadata and its management. *Geomatica* 54: 451–62

Lehmann J, Isele R, Jakob M, Jentzsch, A, Kontokostas D, Mendes P N, Hellmann S, Morsey M, van Kleef P, Auer S, and Bizer C 2014 DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6: 165–97

Li W, Goodchild M F, and Raskin R 2014 Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth* 7: 17–37

Li W, Yang C, Nebert D, Raskin R, Houser P, Wu H, and Li Z 2011 Semantic-based web service discovery and chaining for building an arctic spatial data infrastructure. *Computers and Geosciences* 37: 1752–62

Lin H, Yu B, Chen Z, Hu Y, Huang Y, Wu J, Wu B, and Ge R 2013 A geospatial web portal for sharing and analyzing greenhouse gas data derived from satellite remote sensing images. *Frontiers of Earth Science* 7: 295–309

Lutz M and Klien E 2006 Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science* 20: 233–60

Maguire D J and Longley P A 2005 The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems* 29: 3–14

Masser I 1999 All shapes and sizes: The first generation of national spatial data infrastructures. *International Journal of Geographical Information Science* 13: 67–84

Masser I 2007 *Building European Spatial Data Infrastructures*. Redlands, CA, Esri Press

Mendes P N, Jakob M, García-Silva A, and Bizer C 2011 DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the Seventh International Conference on Semantic Systems*, Graz, Austria

Nogueras-Iso J, Zarazaga-Soria F J, Lacasta J, Béjar R, and Muro-Medrano P R 2004 Metadata standard interoperability: Application in the geographic information domain. *Computers, Environment and Urban Systems* 28: 611–34

Nowak J and Craglia M 2006 *INSPIRE Metadata Survey Results*. Ispra, Italy, European Commission Joint Research Centre

Onsrud H J, Campbell J, and van Loenen B 2010 Towards voluntary interoperable open access licenses for the Global Earth Observation System of Systems (GEOSS). *International Journal of Spatial Data Infrastructure Research* 5: 194–215

Ostensen O M and Smits P C 2002 ISO/TC211: Standardisation of geographic information and geo-informatics. In *Proceeding of the International Geoscience and Remote Sensing Symposium (IGARSS'02)*, Toronto, Ontario: 261–63

Ramage D, Hall D, Nallapati R, and Manning C D 2009 Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore: 248–56

Rennie J D 2001 Improving Multi-class Text Classification with Naive Bayes. Unpublished PhD Dissertation, Massachusetts Institute of Technology

Rose L 2004 *Geospatial Portal Reference Architecture: A Community Guide to Implementing Standards-based Geospatial Portals*. Wayland, MA, Open Geospatial Consortium Discussion Paper No. 04–039

Sebastiani F 2002 Machine learning in automated text categorization. *ACM Computing Surveys* 34(1): 1–47

Tait M G 2005 Implementing geoportals: Applications of distributed GIS. *Computers, Environment and Urban Systems* 29: 33–47

Tang W and Selwood J 2005 *Spatial Portals: Gateways to Geographic Information*. Redlands, CA, Esri Press

Walker W and Maidment D 2006 *Geodatabase Design for FEMA Flood Hazard Studies*. Austin, TX, University of Texas Center for Research on Water Resources Report No. 06–10

Wiegand N and Garcıa C 2007 A task-based ontology approach to automate geospatial data retrieval. *Transactions in GIS* 11: 355–76

Zhang C, Zhao T, Li W, and Osleeb J P 2010 Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web. *International Journal of Geographical Information Science* 24: 903–23

Zhao T, Zhang C, Anselin L, Li W, and Chen K 2014 A parallel approach for improving geo-parql query performance. *International Journal of Digital Earth* 8: 383–402

