Annals of the American Association of Geographers

ISSN: 2469-4452 (Print) 2469-4460 (Online) Journal homepage: http://www.tandfonline.com/loi/raag21

Lowering the Barriers for Accessing Distributed Geospatial Big Data to Advance Spatial Data Science: The PolarHub Solution

Wenwen Li

To cite this article: Wenwen Li (2017): Lowering the Barriers for Accessing Distributed Geospatial Big Data to Advance Spatial Data Science: The PolarHub Solution, Annals of the American Association of Geographers, DOI: 10.1080/24694452.2017.1373625

To link to this article: http://dx.doi.org/10.1080/24694452.2017.1373625

Published online: 14 Nov 2017.

Lowering the Barriers for Accessing Distributed Geospatial Big Data to Advance Spatial Data Science: The PolarHub Solution

Wenwen Li

School of Geographical Sciences and Urban Planning, Arizona State University

Data is the crux of science. The widespread availability of big data today is of particular importance for fostering new forms of geospatial innovation. This article reports a state-of-the-art solution that addresses a key cyberinfrastructure research problem—providing ready access to big, distributed geospatial data resources on the Web. I first formulate this data access problem and introduce its indispensable elements, including identifying the cyberlocation, space and time coverage, theme, and quality of the data set. I then propose strategies to tackle each data access issue and make the data more discoverable and usable for geospatial data users and decision makers. Among these strategies is large-scale Web crawling as a key technique to support automatic collection of online geospatial data that are highly distributed, intrinsically heterogeneous, and known to be dynamic. To better understand the content and scientific meanings of the data, methods including space–time filtering, ontology-based thematic classification, and service quality evaluation are incorporated. To serve a broad scientific user community, these techniques are integrated into an operational data crawling system, PolarHub, which is also an important cyberinfrastructure building block to support effective data discovery. A series of experiments was conducted to demonstrate the outstanding performance of the PolarHub system. This work seems to contribute significantly in building the theoretical and methodological foundation for data-driven geography and the emerging spatial data science. *Key Words: cyberinfrastructure, geospatial big data, semantic classification, spatial data science, Web crawling.*

数据是科学的关键。在今日, 大数据的广泛可及性, 对于促进崭新的地理空间创新形式而言特别重要。本文报导一个应对关键信息基础建设研究问题的最新解决方法——在互联网上提供大型且分散的地理空间数据资源的管道。我首先阐述此一数据取得管道的问题, 并引介其不可或缺的元素, 包含指认信息位置、时空聚合、主题, 以及数据集的质量。我接着提出应对各个数据管道问题、并且让地理空间数据使用者与决策者更容易发现与使用数据的策略。这些策略以大规模网络抓取作为支持自动搜集高度分散、本质上异质且动态的网上地理空间数据之关键技术。为了更佳理解数据的内容与科学意义, 纳入包含时空筛选、以本体为基础的主题分类, 以及服务品质评估等方法。为了服务广泛的科技使用者社群, 这些技术被整合进操作式的数据抓取系统 “极地枢纽” (PolarHub), 该系统同时是支持有效的数据挖掘的信息基础建设的重要基石。本研究进行一系列的实验, 证实 PolarHub 系统的杰出表现。该工作似乎对数据驱动的地理和浮现中的空间数据科学建立理论与方法论基础, 做出显着的贡献。 *关键词： 信息基础建设, 地理空间大数据, 语义分类, 空间数据科学, 网络抓取。*

Los datos son el elemento esencial de la ciencia. La disponibilidad generalizada de *big data* en la actualidad tiene particular importancia para el fomento de nuevas formas de innovación geoespacial. En este artículo se reporta una solución de vanguardia que aboca un problema de investigación clave de ciberinfraestructura—proveyendo acceso expedito a vastos recursos de datos geoespaciales distribuidos en la Web. Primero que todo formulo este problema de acceso a los datos y presento sus elementos indispensables, incluso identificando la ciberlocalización, la cobertura de espacio y tiempo, el tema y la calidad del conjunto de datos. Luego, propongo las estrategias para encarar el asunto individualizado del acceso a lo datos y de hacerlos más fáciles de recuperar, y más utilizables para los usuarios de información geoespacial y para los tomadores de decisiones. Entre estas estrategias se encuentra el rastreo de la Web a gran escala como técnica clave para apoyar la recolección automática de datos geoespaciales en red que se hallan muy distribuidos, son intrínsecamente heterogéneos y que se sabe son dinámicos. Para entender mejor el contenido y significados científicos de los datos, se incorporaron métodos que incluyen el filtrado espacio–temporal, la clasificación temática basada en la ontología y el servicio de evaluación de la calidad. Para servir a una amplia comunidad de usuarios científicos, estas técnicas se integraron en un sistema operacional de rastreo de datos, el PolarHub, que también es un paquete importante de construcción de ciberinfraestructura para ayudar al efectivo hallazgo de datos. Se llevó a cabo una serie de

Annals of the American Association of Geographers, 0(0) 2017, pp. 1–21 © 2017 by American Association of Geographers
Initial submission, July 2016; revised submission, May 2017; final acceptance, June 2017
Published by Taylor & Francis, LLC.

experimentos para demostrar el sobresaliente desempeño del sistema PolarHub. Este trabajo puede contribuir significativamente a edificar los fundamentos teóricos y metodológicos de la geografía orientada por datos y a la emergente ciencia de los datos espaciales. *Palabras clave: ciberinfraestructura, big data geoespaciales, clasificación semántica, ciencia de los datos espaciales, rastreo de la Web.*

Data is the crux of science (Tenopir et al. 2011). The widespread availability of scientific data today has brought researchers into a world in which research is shifting from application driven to data driven (Bell, Hey, and Szalay 2009). Foster (2005) stressed that data are valuable only if others can discover, access, and make sense of it. Data-driven discovery was later proposed by Hey and his colleagues as the fourth scientific paradigm to complement existing paradigms of theory, experimentation, and computation (Hey, Tansley, and Tolle 2009). In the realm of geography, Miller and Goodchild (2015) described the emergence of an evolutionary field, data-driven geography, for the first time in response to rapidly exploding geospatial data. In human mobility research, for example, knowledge discovery is increasingly relying on mining from massive amounts of data collected through geosensor networks, location-based devices (i.e., smartphones), and point-of-sale (PoS) databases (Miller 2010; Kwan 2016).

In health geography, more studies are being conducted using georeferenced social media data to understand public health issues, such as the prevalence of healthy and unhealthy food on a national scale (Widener and Li 2014). This scale cannot be readily accomplished using traditional data collection methods (i.e., questionnaires). In polar science, for instance, researchers are now capable of understanding historical climate change and the impact of global warming on Arctic ecosystems by taking advantage of big data acquired by Earth observation satellites and numerical simulation models (Yin et al. 2011).

Despite the opportunities presented by big data, the deluge of information poses significant challenges to geospatial researchers. One challenge is to identify the best available data distributed on the Web to perform precise science. This data access problem has attracted much attention from various geography-related disciplines (Ramamurthy 2006; Gold 2007; Li et al. 2011; Michener et al. 2011; Allard 2012; Ames et al. 2012). Several U.S. government agencies have acknowledged big data and data access as crucial to future developments as well (Whitehouse 2012). In 2000, the National Science Foundation (NSF) emphasized the importance of "improv[ing] and extend[ing] facilities to collect and analyze data on local, regional, and global spatial scales and appropriate temporal scales" to advance the understanding and prediction of the Earth's environment and habitability (Avery 2000). In the NSF's 2003 blue ribbon report on cyberinfrastructure (CI), data access is identified as one of the four major research themes for revolutionizing science and engineering (Atkins et al. 2003).

Among the emerging CI platforms is PolarHub, a large-scale geospatial data crawler and content analyzer that lowers the barriers for data access across multiple geospatial disciplines developed by the authors. PolarHub integrates large-scale Web crawling, semantic and location analysis, and quality evaluation to tackle the data access challenge in a comprehensive manner. It has the ability to automatically collect distributed data sets, which increases their value for reuse and advances science. It serves as an excellent testbed for scientists to find data to accelerate knowledge discovery process, for monitoring the evolution of service-based data sharing and interoperability, for experimenting with new cyberinfrastructure algorithms to advance spatial data science, and for opening up ample opportunities for applications in multiple disciplines, within and beyond geography.

The rest of the article is organized as follows: The first section formally defines the data access problem and its five indispensable facets. Section 3 discusses the design principles of PolarHub. I then introduce key techniques to identify the theme, spatial and temporal coverage, and quality of crawled data sets. After that, I introduce the graphical user interface (GUI), which integrates proposed techniques, and then conclude the work and propose future research directions.

Distributed Data Access: A Problem Statement

As stated previously, data are the crux of scientific research. Today's big data deluge is affecting the way we do science in nearly all aspects (Hey, Tansley, and Tolle 2009). This is especially true in GIScience, where huge amounts of spatiotemporal data are becoming central in analyzing physical and societal changes. Therefore, making data available to scientists along the knowledge

creation pipeline is always the essential first step. I term this problem a data access problem, defining it in quintuple form to highlight the five key facets for retrieving needed data to support scientific analysis:

$$< \text{cyber} - \text{location},\ \text{theme},\ \text{spatial extent},\ \text{temporal extent},\ \text{quality} > \quad (1)$$

The first facet is about the Web location where data sets are shared and hosted, as scientists are always most concerned about "where to find data." Today's paradigm for data discovery is very different from the traditional approach of copying data. The development of the Internet has made geospatial data easily shared and widely distributed. The exponential growth in the amount of information that the Web carries, however, presents significant challenges in locating relevant spatial data as it accounts for only a very small percentage of the entire volume of data on the Web. Therefore, identifying the footprint and entry point of Web-based spatial and spatiotemporal data becomes one of the most important dimensions to improving data accessibility.

The second facet is to identify the theme of a data set such that it can be properly used to support an analysis. Geospatial data sources have become much more diverse due to the development of Earth observation programs and the advancement of simulation models. As a result, semantically distinguishing the content of a data set has become a challenge (Halevy 2005).

The third and fourth facets are about the spatial and temporal coverage of the data sets. Geophysical and social phenomena differ from locality to locality and might present strong space–time variability due to local terrain, human activity, and land use patterns (New, Hulme, and Jones 2000). Moreover, geospatial research most often covers study areas, even related ones, using different parameters. Therefore, the data need to not only be relevant but cover specific geographical area(s) and period(s) of time to enable scientific analysis.

The final facet is about data quality or the degree of uncertainty or reliability that is acceptable to solve a research problem at a certain scale. Bad or unreliable data will result in errors, making the information useless for interpreting the dynamics of various physical or social phenomena. Hence, an effective data discovery tool must not only find the data set but it must also be able to evaluate the data quality to deliver the best, most reliable data to the scientists.

Each of the five facets is therefore indispensable in ensuring a successful spatial data access and retrieval process. No matter how smart a cybertool is, there might still be a lack of key data to analyze if the cyber-location (or accessibility) challenge is not addressed. Similarly, a data set, even though available, might not effectively support research if spatial, thematic, and quality issues (semantic challenges in short) are not taken into consideration.

Literature

In this section, I review efforts in addressing both the accessibility challenge and semantic challenge in the literature.

Data Access through the Spatial Data Infrastructure

To increase public access of distributed geospatial data services, various national and international spatial data infrastructure (SDI) research has been initiated. Exemplar solutions include the U.S. Government's open data portal, data.gov (Lakhani, Austin, and Yi 2010), the European Union's INSPIRE project (EU INSPIRE 2007), and the Global Earth Observation System of Systems (GEOSS; Christian 2005). These SDI solutions advance data access and discovery using a Web catalog, which data providers register to be included in and publish their data products into. Data users are then able to search the catalog for data sets that match their space and time interests. These solutions suffer from significant limitations, however, including data collection that relies heavily on voluntary data submission (Li et al. 2011) and outdated data. For example, once metadata are registered in the catalog, there is often not an effective mechanism to inspect or update them as the data evolves. As a result, dead Web links are common due to coverage, availability, or Web location changes (Li, Yang, and Yang 2010).

Overcoming this issue has fostered the development of cross-catalog harvesting (Li, Yang, and Raskin 2009), which creates a data-sharing channel among SDIs. Although this has helped identify some data, data discovery is still restricted within the scope of a limited number of known catalogs. Other solutions such as catalog search engines that search for online catalogs and then make data discovery through catalog harvesting have been developed, but they commonly fail to find individual services not part of any catalog. In fact, as the Web expands, online geospatial data resources are

expected to become more distributed rather than centralized, making this a significant issue. Thus, a catalog solution without the ability to realize automatic data discovery will be nearly impossible to sustain.

Active Data Discovery through Web Crawling

Another cluster of efforts for active data discovery centers on collecting geospatial data from social media, known as volunteered geographic information (VGI; Goodchild 2007). These crawlers (Gao et al. 2014; Widener and Li 2014; Wang et al. 2015) use the customized application programming interface (API) provided by a specific social media Web site instead of looking for data from the unstructured Web. The challenges in terms of (1) handling the diversity and complexity of unstructured data in the deep Web and (2) effective extraction of spatially annotated data are therefore much lower than developing a new tool set to provide widespread access to data distributed on the Web.

Lately, geospatial researchers have attempted to employ Web crawling techniques for active discovery of geospatial data and services (Li, Yang, and Yang 2010). This has presented vast challenges such as big data storage and management, limitations in computing power, and sophistication in crawler designs. As a result, most existing solutions adopt a strategy of metacrawling (Lopez-Pellicer et al. 2011), a way to filter data of interest from search results in commercial search engines, such as Google or Bing (Huang and Chang 2016). Although the algorithm design is simplified, coverage of the Web is questionable, especially for the geospatial Web, due to its heavy dependency on another search engine. Moreover, few of these tools provide a thorough analysis of the collected data, impeding their comprehension and usability.

Semantic Enhancement for Data Discovery and Understanding

Besides increasing access, a comprehensive data retrieval engine should promote data comprehension to enhance reuse. A key aspect is for semantic understanding of metadata to facilitate topical classification and search (Sicilia 2006). It is widely acknowledged that different information communities frequently use the same keywords to refer to different objects or phenomena or use different keywords to refer to the same phenomenon (Worboys and Dean 1991; Bishr 1998; Fonseca et al. 2002; Li, Yang, and Raskin 2008; Goodchild and Janelle 2010). Even within the same community, the use of keywords for describing the same entity might change over time (Ventrone 1991). These discrepancies result in a lack of understanding of the data content or semantic heterogeneity (Lutz et al. 2009), which significantly impedes the usability of data. This challenge is common for both topical keywords and those involving place names (Li, Goodchild, and Raskin 2014).

Addressing this issue always involves the use of ontology, a machine-understandable knowledge base for semantic annotation and concept mapping (Bukhres et al. 2000; Raskin and Pan 2005; Li, Yang, and Raskin 2008; Janowicz and Hilzler 2013). However, the effectiveness of this approach depends heavily on the completeness of the ontology in use (Li, Goodchild, and Raskin 2014). The population of ontologies is always time consuming and sometimes debatable across domains and knowledge areas (Agarwal 2005). A more intelligent and automatic solution is therefore desired.

In this article, I propose and develop cutting-edge solutions to tackle these data access challenges in a comprehensive manner and integrate them into an operational cyberinfrastructure platform for Web-wide access.

PolarHub Design Principles

The PolarHub cyberinfrastructure platform addresses questions in the form of "Find me some online data set (U to denote its Web location) related to theme A within a study area of X from Time Y_1 to Y_2 at a quality of or better than Q." The variables—U, A, X, Y_1, Y_2, and Q—match the five facets for accessing needed data sets identified in the quintuple form. There are two ways of approaching this question. One is to develop a theme-specific crawler. The other is to build a crawler that tries to find all possible geospatial services and then classifies them into different themes and different coverage regions as needed. The advantage of a theme-specific crawler is that it only collects the data that address the exact needs of a specific application or domain, such as polar climate science. Because irrelevant data sets are filtered out on the fly, however, when geospatial data related to another theme are needed, the crawling process must be repeated with new requirements, resulting in a waste of computing resources and additional time.

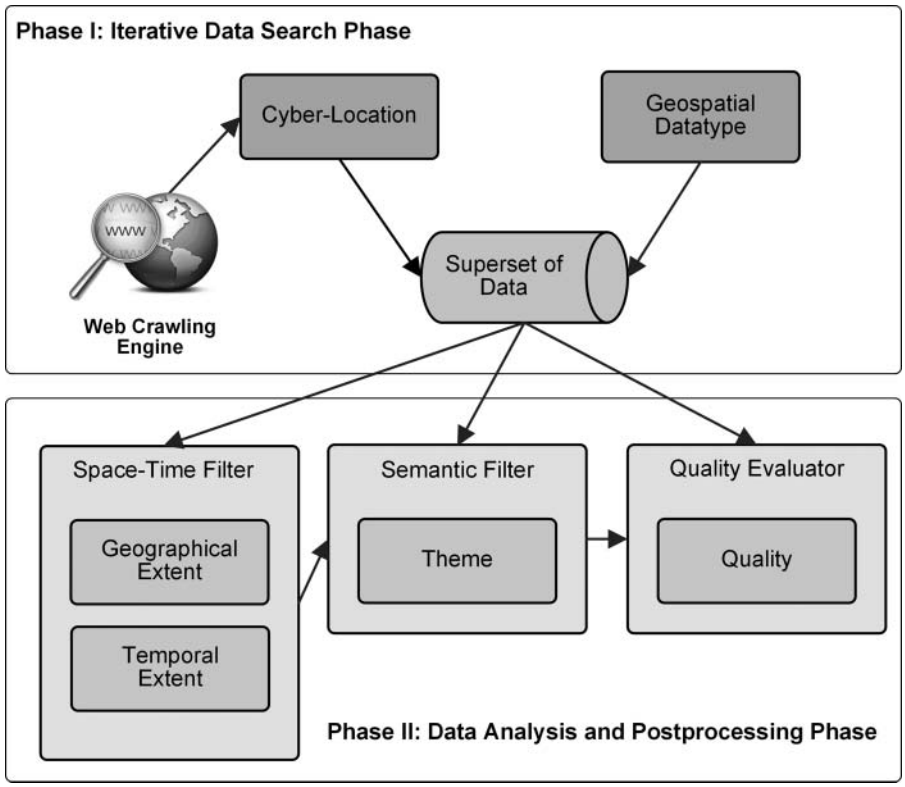

Figure 1. PolarHub's multiphase data retrieval and search workflow.

The design of PolarHub uses the second principle to collect as many geospatial data services as possible through large-scale Web crawling regardless of the data's theme or geographical extent. Once found, the data are further analyzed and categorized (Li, Wang, and Bhatia 2016). In this way, the complex data access problem is decomposed and tackled at different phases. Figure 1 demonstrates PolarHub's data retrieval and search workflow.

The first workflow phase identifies all possible geospatial data that exist on the Web regardless of their theme, spatial or temporal extent, and quality through a continuous Web crawling process. The goal is to collect as many data sets as possible. PolarHub discriminates geospatial data from other domain data based on data types. For instance, to foster the widespread sharing and interoperability of geospatial resources, geospatial data are often encapsulated into standard data services compliant with Open Geospatial Consortium (OGC) standards.

A geospatial data set can be published into an OGC Web Map Service (WMS; de La Beaujardiere 2006) that renders the actual data into maps. Vector data can be serialized and shared through an OGC Web Feature Service (WFS; Vretanos 2005). A raster data piece, on the other hand, can be published according to the OGC Web Coverage Service (WCS; Whiteside and Evans 2008) to deliver the coverage data in a standard format, such as GeoTiff (Ritter and Ruth 2000) or ArcGrid (ESRI 2011). There are also other service standards for sharing sensor observation data, such as the OGC Sensor Observation Service (SOS; Na and Priest 2007) or sharing tiled geospatial data such as the OGC Web Map Tile Service (WMTS; Masó, Pomakis, and Julià 2010). These data sets, available as services, have specific patterns that can guide the data search and extraction process. The way PolarHub works to find these data is discussed in detail later in this article.

In parallel with the search process, a postprocessing program is initialized to analyze collected data sets and pull out a subset that satisfies the needs of a domain application. Three filters, a space–time filter, semantic filter, and quality evaluator, are applied. The space–time filter returns data that depict information about the study area, such as the Arctic, at the predefined time period. The semantic filter classifies the data set according to topics or keywords to obtain a subset that contains relevant themes; that is, the desired geophysical, atmospheric, or socioeconomic phenomena. I

name this module a *semantic filter* to emphasize the importance of data semantics in ensuring an accurate thematic classification of the data set. The third module is the quality evaluator. This module diagnoses the availability of data as well as its service performance by measuring the response time, server stability, and data download success rate, among other factors. The quality evaluator is essential in delivering the most stable data to end users.

These three modules can run sequentially or concurrently because there is no interdependency among these modules. The successful execution of these modules, though, relies on the analysis of the metadata from the data services. A *GetCapabilities* request, supported by all OGC data services according to the definition in OGC Common (Whiteside 2007), enhances the compatibility and communication among distributed data services. In this way, new data sets and data service metadata are acquired and saved in PolarHub's local data repository between the data search phase and the postprocessing phase to support various filtering operations.

Key Techniques to Enable Effective Access to Distributed Geospatial Data

Where Are the Geospatial Data Located?

PolarHub uses a hybrid search approach that combines Web crawling and a metasearch strategy to achieve high-performance crawling. The two strategies (crawling and metasearch) do not appear to be harmonizable. In this specific search problem, however, the coordination of the two strategies realizes high crawling efficiency. The metasearch takes advantage of the huge index of general search engines to narrow the search scope to geospatial-related Web content of interest. The Web crawling then spreads the search out from these Web seeds to discover more data and service resources. This strategy avoids aimless crawling because the metasearch eliminates many irrelevant Web pages. Figure 2 demonstrates PolarHub's software architecture design. The core search engine is where the hybrid search strategy is implemented. Initially,

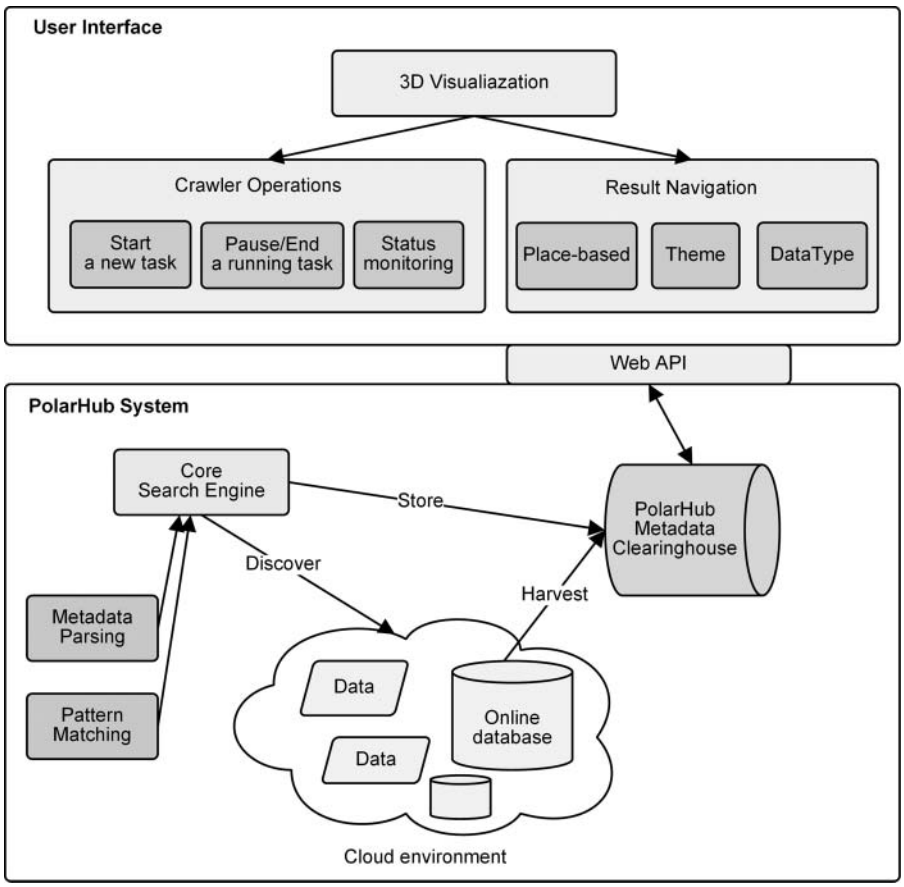

Figure 2. PolarHub architecture design.

keywords with either a thematic topic (i.e., land use), an agency name (i.e., NASA), a place name (i.e., Greenland), or any combination are redirected to general search engines, such as Google or Bing, to start the metasearch. Note that the data or service type of interest, such as Web Map Service, is jointly used with these keywords for further Web content filtering. Once initial search results are retrieved from these search engines, overlapping Web pages are removed and the rest are used as the crawling seeds to start general crawling. General crawling iteratively retrieves the source Web page of a seed Web page, extracts all of the hyperlinks from it, and then visits the linked Web pages until a given crawling depth, measured by the number of jumps from seed Web page to current Web page, is reached. Concurrently executed with the crawling process, a metadata parsing module is enabled to facilitate the determination of whether a URL belongs to an endpoint data set or a data service by matching service patterns encoded in regular expressions (Li, Wang, and Bhatia 2016).

Besides the combined crawling strategy, another unique feature of PolarHub is its ability to search for both geospatial data services and the online databases that host these services. This feature is of great importance because it integrates the advantages of (1) a data crawler, which focuses on searching for scattered online data sets that are not registered in any catalog (Li, Yang, and Yang 2010); (2) the centralized catalog, which conducts cross-harvesting of data sets residing in other known catalogs (Li, Yang, and Raskin 2009); and (3) a catalog crawler, such as Spatineo Directorate, which searches for distributed catalogs in an active manner and then retrieves geospatial data from discovered catalogs.

Using this strategy, PolarHub can mine and discover distributed geospatial data sets from both the surface Web for scattered data sets and the deep Web (Bergman 2001) for data hidden within a database using a Catalog Service for the Web (CSW) interface. This significantly improves the discoverability of distributed geospatial data and services.

What Is the Theme of Each Data Set?

Annotating the data set by its theme or subject is a typical classification or categorization process (Biettron, Pallu, and Tricot 2006). In the context of this research, thematic classification is performed on the metadata (textual information) describing the content of the data set. This is different from thematic classification in remote sensing, which classifies satellite imagery to obtain the land cover types (Rosenfield and Fitzpatrick-Lins 1986; Vatsavai and Bhaduri 2011).

Metadata thematic classification is of great importance to geospatial applications. First, almost without exception, geospatial applications must support mapping and the generation of maps showing the environmental or socioeconomic conditions of a geographical area. These maps are often called thematic maps (Scholl and Voisard 1990), as they overlay geospatial data with different themes to support geospatial knowledge acquisition (MacEachren 1991). Organizing PolarHub-identified data sets thematically greatly facilitates the creation of thematic maps.

Second, annotation according to the data set's theme is an effective way to index data and avoid information overload during the retrieval and search process, especially within a large collection of data (Clark and Watt 2007). Existing SDI systems, such GEOSS (Bai et al. 2012), the Global Change Master Directory portal (GCMD; Miled et al. 2011), and Data.gov (Lakhani, Austin, and Yi 2010), rely on manual selection of data themes by data providers. When dealing with massive data sets, this method is not efficient. In PolarHub, a methodology that combines ontology and advanced metadata processing is used to realize automated thematic classification of crawled data sets. Figure 3 illustrates the proposed thematic classification framework. Figures 4 and 5 provide examples of input data and knowledge for thematic classification, respectively.

First, the structured metadata (green box) is parsed. Most metadata records are encoded in popular formats such as XML and JavaScript Object Notation (JSON). Figure 4 demonstrates an XML fragment describing one of the data layers contained in a U.S. Geological Survey (USGS) data set used as input data. In addition to these commonly used metadata standards, the OGC community is investigating the use of Semantic Web standards, such as Resource Description Framework (RDF) or Web Ontology Language (OWL), to add semantic tags for machine understanding.

These metadata are then sent to the text-processing component, in which the description information about the data set included in the "Abstract," "Title," or "Keyword" tags are extracted. This textual information is further processed to extract keywords using N-gram analysis (Bespalov et al. 2011). The open

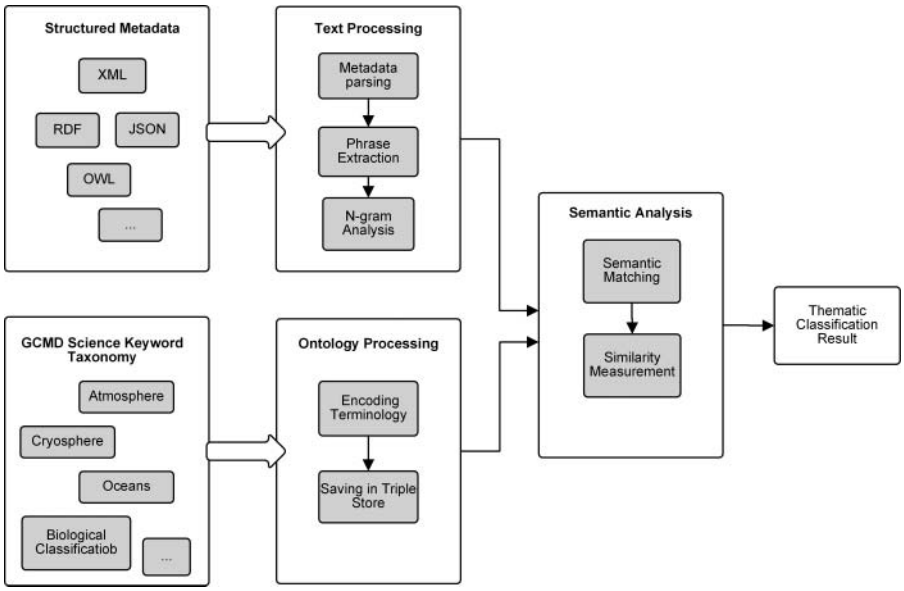

Figure 3. Thematic classification workflow. *Note:* XML = Extensible Markup Language; RDF = Resource Description Framework; JSON JavaScript Object Notation; OWL = Web Ontology Language; GCMD = Global Change Master Directory.

source n-gram analyzer NgramTool (Naqao and Mori 1994; Lu, Zhang, and Hu 2004) is used to parse metadata in PolarHub.

The n-gram analysis is adopted to detect and extract cooccurring words within a predefined window. In the context of this work, the scanning window is limited to be within each sentence. The proposed N-gram analysis breaks a sentence or phrase into sets of words. For example, N = 2 means each pair of consecutive words in the sentence (first/second word, second/third word, third/fourth word, etc.) were extracted. These extracted phrases are known as bigrams. Trigrams or multigrams can also be generated by extracting three consecutive words (N = 3) or any number of grams

```
▼<Layer queryable="0" opaque="0" cascaded="0">
   <Name>Top5_Li</Name>
   <Title>Lithium in the Top5 horizon</Title>
   <Abstract>Data calculated from DS-801</Abstract>
  ▼<KeywordList>
   ▼<Keyword>
      geoscientificInformation; soil horizons;
      chemical analysis; Lithium
    </Keyword>
   </KeywordList>
   <LatLonBoundingBox minx="-165" miny="24" maxx="-66"
   maxy="73"/>
   <BoundingBox SRS="EPSG:4267" minx="-165"
   miny="23.9995" maxx="-66.0005" maxy="73.0006"/>
  ▼<Attribution>
   ▼<Title>
      Geochemical and mineralogical data from soils in
      the conterminous United States
    </Title>
    <OnlineResource
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xlink:href="http://mrdata.usgs.gov/ds-801/"/>
   </Attribution>
```

Figure 4. Example of input data: Metadata fragment in XML.

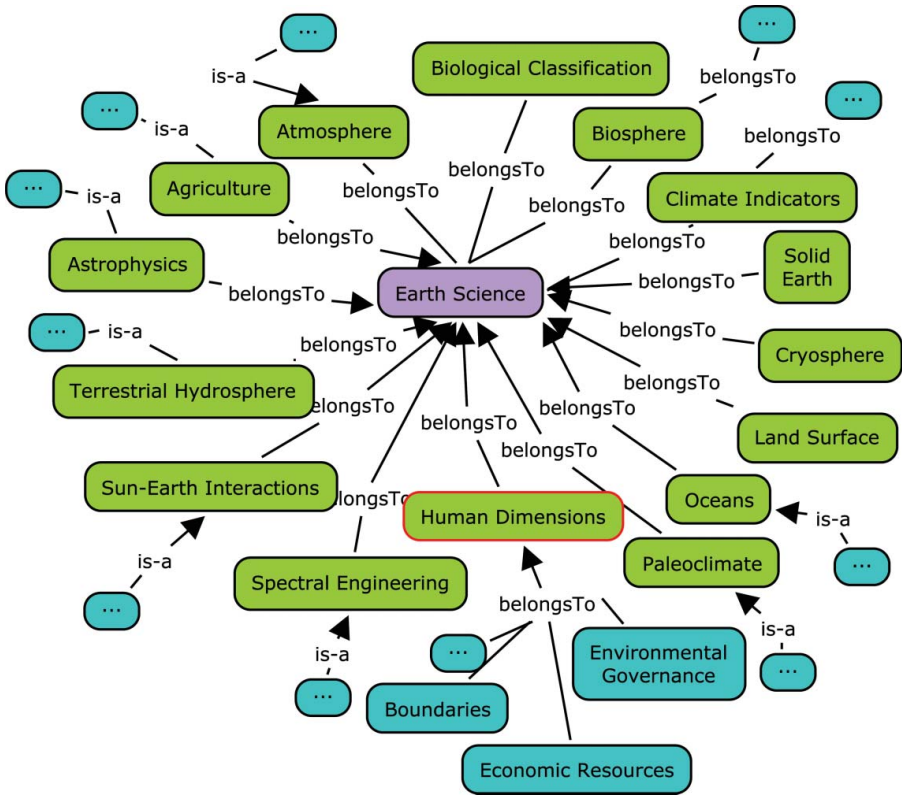

Figure 5. Example of knowledge for thematic classification: A snapshot of Global Change Master Directory science keyword structure. (Color figure available online.)

(four grams, five grams, etc.), as determined by length of the original phrase. This preserves the complete semantic information by considering a phrase rather than just a single word in the semantic analysis process, ensuring more complete thematic classification.

After text processing, all scientific keywords and phrases are saved in a document vector for cross-comparison with knowledge defined in the ontology. In this work, the GCMD science keyword taxonomy is adopted. There are more than 3,000 keywords in total within the taxonomy, making it one of the most comprehensive knowledge bases for Earth and space science topics. As Figure 5 demonstrates, the Earth science domain is divided into fifteen subdisciplines (e.g., land surface, atmosphere, etc.). The keywords are then populated to domain ontology, encoded using RDF format, and saved in a triple store.

After this preprocessing, the set of keywords ($m_i \in M$) extracted from the metadata and the set of theme keywords ($t_j \in T_j$) defined in the ontology are sent to the semantic analysis module for keyword matching. Rather than matching only the appearance of the keywords, a semantic matching process is introduced. That is, the relevancy of metadata keyword m_i and theme keyword t_j is measured by how similar each pair of words is semantically. Normalized Google Distance (NGD) defines this similarity, namely,

$$NGD(m_i, t_j) = \frac{\max\{\log f(m_i), \log f(t_j)\} - \log f(m_i, t_j)}{\log K - \min(\log f(m_i), \log f(t_j))}, \quad (2)$$

where K is the total number of Web pages searched by Google, and $f(m_i)$ and $f(t_j)$ are the number of hits when the metadata keyword and the theme keyword are used for searching, respectively. $f(m_i, t_j)$ is the number of Web pages in which both m_i and t_j appear. To ensure high accuracy in the thematic classification process, a high NGD threshold (0.8) was set for determining whether a metadata belongs to a certain theme. If more than one category receives a score higher than the given threshold, the data layer is assigned to the category that gets the highest score. During the semantic analysis and matching process, terminologies from another Earth science ontology, the Semantic Web for Earth and Environmental Terminology (SWEET; Raskin and Pan 2005), are

integrated to enrich the existing knowledge base. Classification results are reported later in this article.

How Are Data That Cover a Specific Period or Geographical Area Distilled?

The next challenge for PolarHub is how to distill data that cover a specific period or geographical area. In other words, a time stamp is needed as well as a spatial filter to help PolarHub identify data sets for searches such as "Find me all data services covering East Europe," or "Find me all polar data services" retrieved in "year 2000."

Extraction of OWS metadata time stamps requires the development of a time parser for XML capability files. Two types of services, the Web Map Service Time (WMS-T) and SOS usually contain time components in their metadata. For the OGC WMS-T, the mandatory field "*wms_timeextent*" defines the valid time extent for a data layer. For SOS, the time parse identifies the "*Time Period*" field that provides temporal coverage of data. Within each individual record, the timestamp at which the data are collected is also provided. These two parameters are the key to extracting temporal properties of a data set and time filtering.

For PolarHub, a regular expression-based pattern analyzer detects different time patterns. In the time filter module, separate time parsers, one for temporal coverage and one for data collection, are developed and integrated as plug-ins. They are dynamically invoked when parsing and filtering of a specific data service is requested.

For distilling geographic area data, a commonly used approach is the comparison of spatial information within the region of interest in a query. This method always involves geocoding, a process that converts a place name into georeferenced spatial data, normally represented by its latitude and longitude (Goodchild 2013). In gazetteers (a geographical dictionary or directory used in conjunction with a map or atlas) like DBpedia (Auer et al. 2007) and GeoNames (Vatant and Wick 2006), the provision of georeferenced place information is through points, which omits boundary information. This information is important, especially when the place is large in extent, such as the state of Maine. In spatial data infrastructure solutions, the spatial filter is always conducted by detection of overlap in the spatial extent that a data set covers and the spatial extent of region of interest (ROI) provided by the user (Maguire and Longley 2005).

Based on several preliminary experiments, however, many data services, particularly those whose sources are remote sensing platforms that provide data covering the whole world, spatial coverage is defined as "−180, −90; 180, 90" in a "latlon" coordinate system. Even though these data are considered relevant for many spatial queries, they are often not detailed

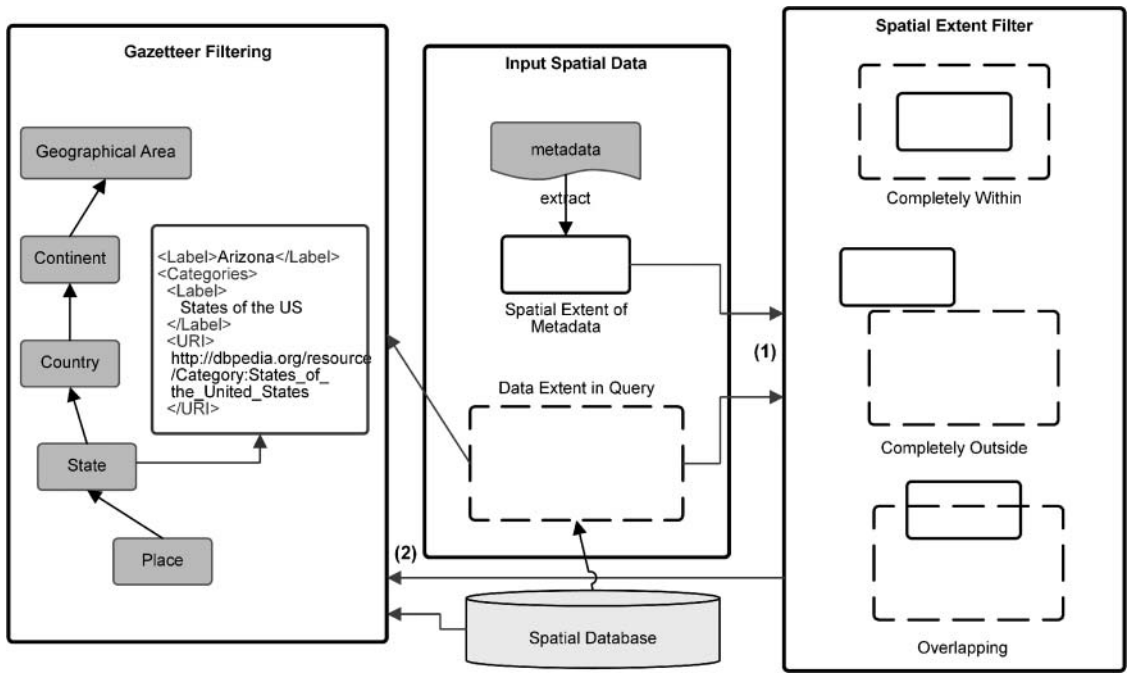

Figure 6. Workflow of the hierarchical spatial filter.

enough to provide information about a specific region of interest. In addition, the coordinate information for some data sets might be missing in its metadata, excluding such data from a search. To address this, I introduce a hierarchical spatial filter that introduces the combined use of spatial information and semantic information in the analysis (Figure 6).

For a data set that claims to cover a world extent but in fact only covers a subregion or a data set with missing latlon information, a gazetteer is integrated to narrow down the actual geographical extent of coverage and improve the query effectiveness. This is accomplished by first extracting the place name existing in the title of the metadata. Next, the Flicker map API is used to obtain the bounding box information of the place. Note that this process might introduce an issue of place-name ambiguity, a common case, as different regions often share the same place name. Ambiguity can also occur when there is more than one place name appearing in the title; for example, "New Zealand percentage change in regional emigration to Australia."

To eliminate ambiguity, two scans were performed. The first scan is on the title field, which my preliminary study showed to be most informative when referring to a data set's coverage area. The second scan is for the appearance of the same place name in other data sets provided by the same Web service or Web host, as most typically provide regional data, especially when the same place name appears in the metadata of multiple data sets. Next, the accurate latlon information of the place name defined in those closely related data sets was used to interpret the actual spatial extent of the data set being examined.

Once the spatial extent (bounding box with solid line in Figure 6) of the input data set and the ROI in a filtering request (bounding box with dashed line in Figure 6) are identified, the spatial extent filter is applied to determine the relatedness of a spatial data to a filtering request. This presents three scenarios that determine whether the data set covers the requested area.

The first scenario is when the coverage of a data set is completely inside the requested spatial extent, as shown in the top case in the "spatial extent filter" component (far right box) of Figure 6. This is determined by all points of the data record (solid rectangle) being within that of the requested spatial extent (dashed rectangle). One can easily see that this data set will provide data covering the requested area of interest. The second scenario is when the two spatial extents have no overlap at all. This data record can be filtered out, as it will not provide data covering the requested spatial extent. The third scenario is when spatial extents overlap but not all data points are within the requested spatial extent. When the two spatial extents have overlap, the following criterion is adopted to determine whether a data record satisfies a spatial filter request:

$$\min\left\{\frac{A(\gamma)\cap A(\delta)}{A(\gamma)}, \frac{A(\gamma)\cap A(\delta)}{A(\delta)}\right\} \geq \theta_1, \qquad (3)$$

where γ, δ represents the spatial extent of a data record and an area of interest in the filtering request and A (γ) and A(δ) denote the area they cover. The threshold θ_1 is set at 25 percent. That means that the overlapping coverage must be larger than 25 percent of the total area covered by a data record and the total area in the spatial filtering request to make sure that the data provided are detailed enough to provide sufficient information.

How Good Is the Quality of the Data?

In addition to space, time, and theme, quality of data services (QoDS) plays a key role in data- and service-sharing infrastructure to help connect the research community with data of interest (Foster 2005; Mani and Nagarajan 2005). QoDS can be further categorized as quality of metadata (QoM) and quality of service (QoS) for data provided over the Internet. QoM focuses on evaluating the quality of data and service metadata used to identify the content of the data service. The main criteria for evaluating QoM is its completeness in providing information such as space, time, and theme (Fox and Hendler 2014). Because metadata are generated during the service curation process, as soon as the service becomes available, QoM is predetermined. It is very much data provider centric rather than user centric. In contrast, QoS provides SDI and geospatial cyberinfrastructure end users a good indication of the reliability and performance of a remote service. Thus, it is very important in the data integration and analysis process. Because this work focuses on providing a user-centric system, I mainly discuss the development of the QoS rather than the QoM in this article.

QoS is very much a subjective process. Different works in the literature propose various quality indicators that evaluate the availability, accessibility, integrity, reliability, security, and other quality-related factors (Mani and Nagarajan 2005). Li et al. (2011) developed a service quality checker to provide

performance information by calculating the response time of *GetCapabilities* request from OGC OWS. They found that response time is the main factor used for the performance measurement. Xia et al. (2015) enhanced the model and developed a complex quality evaluation framework that considers the location of the data server and location of the users at different time periods, such as different days of the week. Interesting results were obtained. For instance, the response time is faster during weekends than weekdays because of the low volume of researchers using the data services. This information requires the simulation of large numbers of users and statistics over a long time period (a year) to obtain quality information, however. For a newer service, this model cannot be applied as a quality score cannot be provided on the fly.

In addition, the Federal Geography Data Committee (FGDC) offers an online service quality evaluator that provides measurement on more than twelve types of spatial data services. Integrating the FGDC quality evaluator with PolarHub enables the discovered services to be evaluated in real time. Two primary performance indicators are adopted—response time and reliability. Response time has the same usage as that proposed in earlier works. Reliability is a weighted combination of the speed of performance (measured by the response time t) over a test period τ (i.e., seven days) and the rate of successful requests.

Assume $m = f(s_i, \tau, n)$ returns the number of successful requests that is tested on service s_i over n tests during the time period of τ. Mathematically, the numerical performance score $QoS(s_i)$ indicating the reliability of service s_i can be expressed as

$$QoS(s_i) = \left(w_1 \times \left(1 - \frac{t}{c}\right) + w_2 \times \frac{f(s_i, \tau, n)}{n} \right) * 100,$$

where w_1 and w_2 are the weights, both with a value range of [0, 1], and $w_1 + w_2 = 1$. c is a constant set as the longest response time by statistics. The value range of $QoS(s_i)$ is between 0 and 100, with 0 unavailable data sets and 100 referring to the service with the highest quality.

Experiments and Results

Scalability of PolarHub in Discovering Geospatial Data Services

PolarHub's finishing condition is controlled by two parameters—crawling depth and width. The crawling depth records the number of clicks it takes to visit a Web page from a seed Web page. The crawling width determines the number of relevant seed Web pages to be included in a crawling process. Figure 7 illustrates

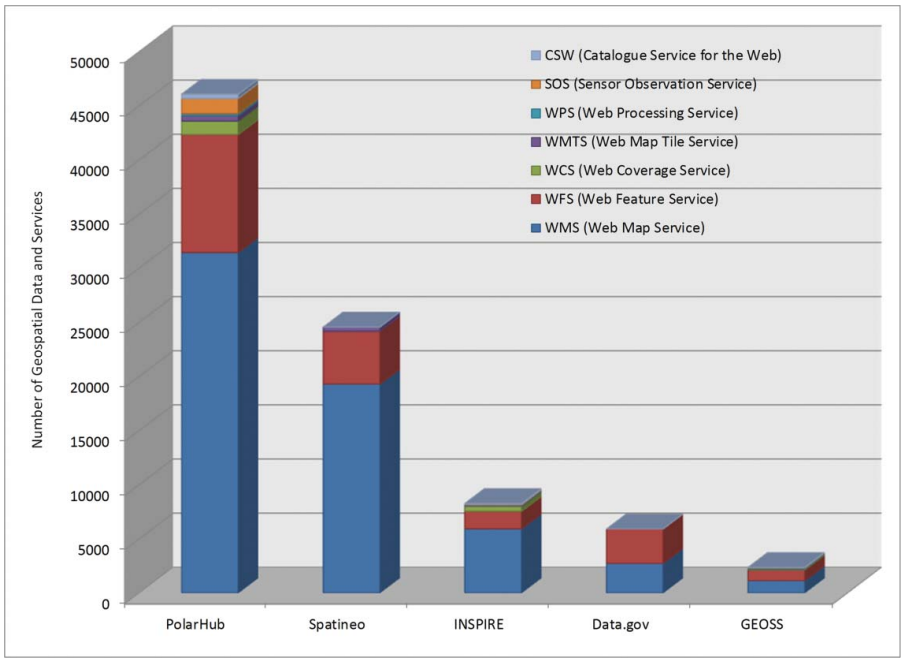

Figure 7. Comparison between PolarHub and other spatial data infrastructure/crawler solutions on the ability to collect distributed geospatial resources. (Color figure available online.)

the number, type, and distribution of PolarHub and other geospatial data service collection systems. These data services were found using more than 100 crawling tasks, which included general keywords containing only the service type, such as "web map service" or "web feature service," as well as thematic keywords extracted from the GCMD science keyword taxonomy.

To date, PolarHub has found almost 77,000 data sets distributed in ninety-five countries. Of these, about 29,000 were found from direct crawling and 16,000 from harvesting of distributed catalogs. This number is significantly higher than other existing solutions, including 5,140 from data.gov, 15,274 from INSPIRE, and 33,314 from Spatineo. Overall, it took PolarHub approximately six months, after initiation of the first prototype, to accumulate these data services. The geographical distribution of these services is illustrated in Figure 8. As shown, the United States, Canada, many European Union countries, and Australia are major sources of geospatial data shared as services. This pattern matches the open government movement initiated by these countries.

In summary, the outstanding performance of PolarHub for data discovery lies in two design advantages. First, it combines metasearch with large-scale crawling strategies. This is significantly different from those relying on manual data registration. Second, PolarHub not only searches for distributed catalogs and conducts further service harvesting from these catalogs but it also searches for scattered geospatial data on the Web. This hierarchical crawling harvesting strategy makes it outperform other solutions, especially those that only crawl for data within catalogs.

Accuracy of Thematic Classification

As noted earlier, thematic organization of PolarHub-identified data sets is of great importance to geospatial applications. Figure 9 further illustrates the thematic classification results of the discovered Earth science data set. Data services related to human dimensions has the highest number, occupying about 35 percent of the entire data collection. The other physical geography topics occupy a cumulative 65 percent. This distribution might reflect the importance of human factors in geospatial and Earth sciences (Reid et al. 2010).

To illustrate the advantages of using semantic analysis to support thematic classification, the differences in three thematic classification approaches were analyzed. The first approach was using no semantics, only

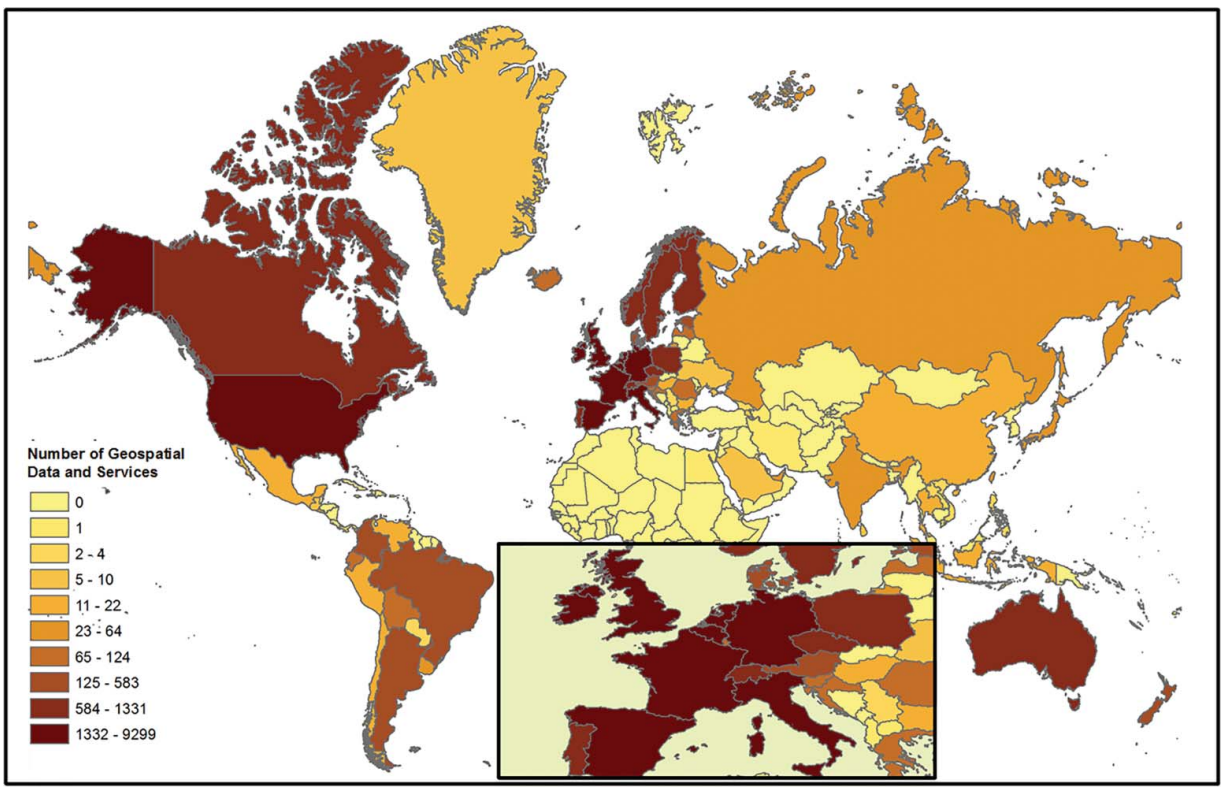

Figure 8. A geographical distribution of PolarHub-identified geospatial services. (Color figure available online.)

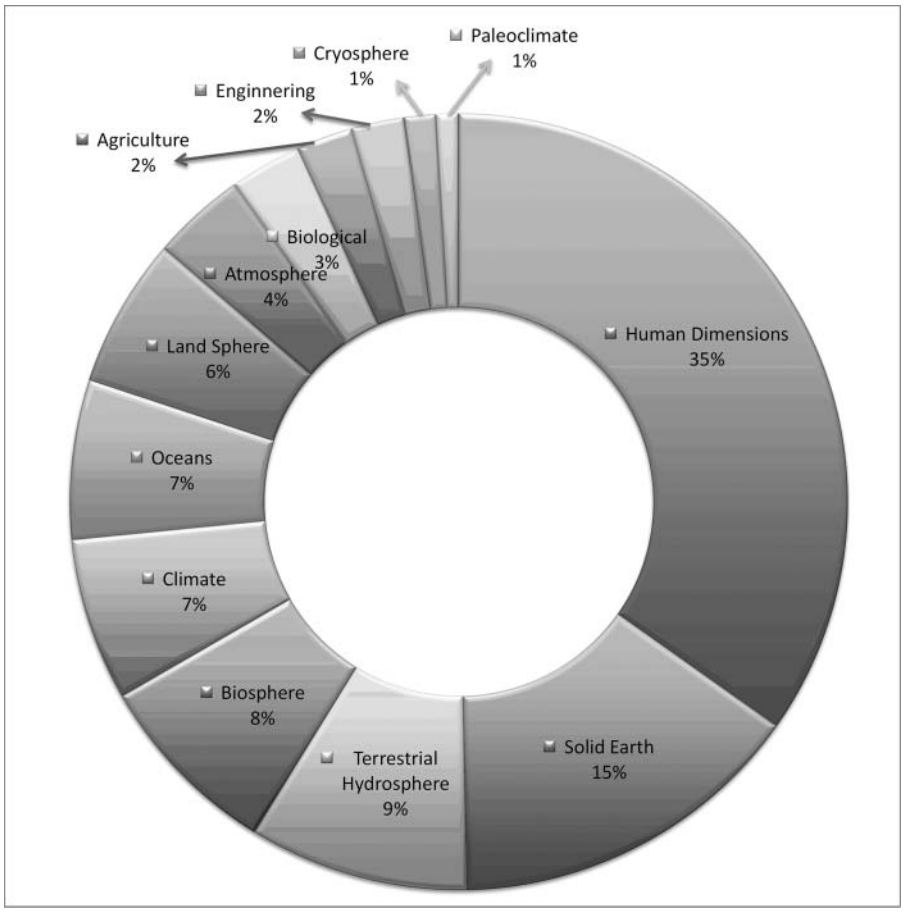

Figure 9. Thematic classification results on each Earth Science topic.

pure key word matching. This approach only matched the words listed in the topic with the metadata records of the OWS. The second approach was GCMD-based thematic classification. In this case, not only the top-level themes were matched, but the subcategories of each theme provided in the GCMD ontology were also matched. The largest depth in the GCMD ontology can reach up to seven. After keyword matching, the results from each subcategory were aggregated and the total number of matched records from all these subcategories were assigned as the final thematic classification results. The final approach was PolarHub's combined use of a taxonomy that integrates scientific terminology from GCMD and SWEET ontology as well as the semantic matching process. Results are demonstrated in Figure 10.

From Figure 10, it can be seen that when only keyword matching is used, few records can be thematically classified. In comparison, the domain GCMD ontology substantially increases the number of total matched records (eighteen times higher than when pure key word matching is used) across all topic areas. When the semantic analysis and classification is applied, 60 percent more data records are successfully classified.

In addition to number of data being successfully classified, experiments were conducted to evaluate the classification accuracy using the advanced semantic-based methods. To make the evaluation manageable, thirty sample data sets were randomly selected from each thematic category. Two accuracy results were generated. One used purely GCMD taxonomy; the other applied the combined taxonomy and text processing approach. Figure 11 illustrates the results.

It can be observed that the GCMD approach achieves a classification accuracy of higher than 80 percent for all categories. On average, the overall accuracy is above 90 percent when it is weighted by the number of data sets belonging to each category. This shows the benefits of introducing scientific taxonomy in the classification procedure. I did note some limitations in using the GCMD taxonomy alone, however. For some categories, it gets lower classification accuracy due to

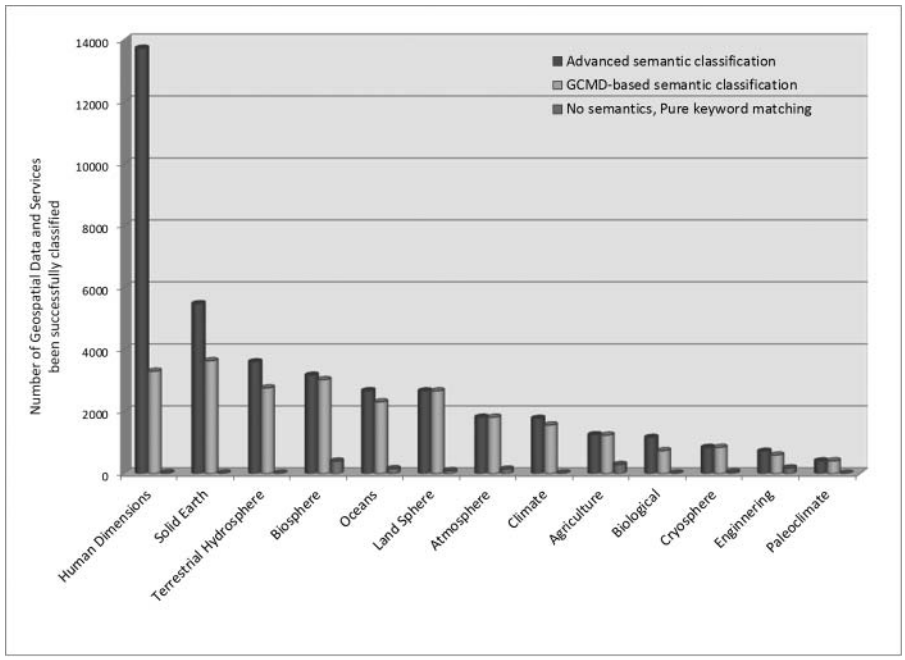

Figure 10. Comparison in classification result before and after semantic analysis is applied. *Note:* GCMD = Global Change Master Directory.

ambiguity. For example, carbon is found in multiple classifications. It is listed as part of the leaf node in Agriculture under the hierarchy Agriculture\Soils\Carbon as well as a component of the atmosphere. This can result in a misclassification of data if the carbon is classified under Atmosphere rather than under Agriculture.

Using the hybrid approach enhances the GCMD approach by adding text processing. The results show improved classification accuracy for four of

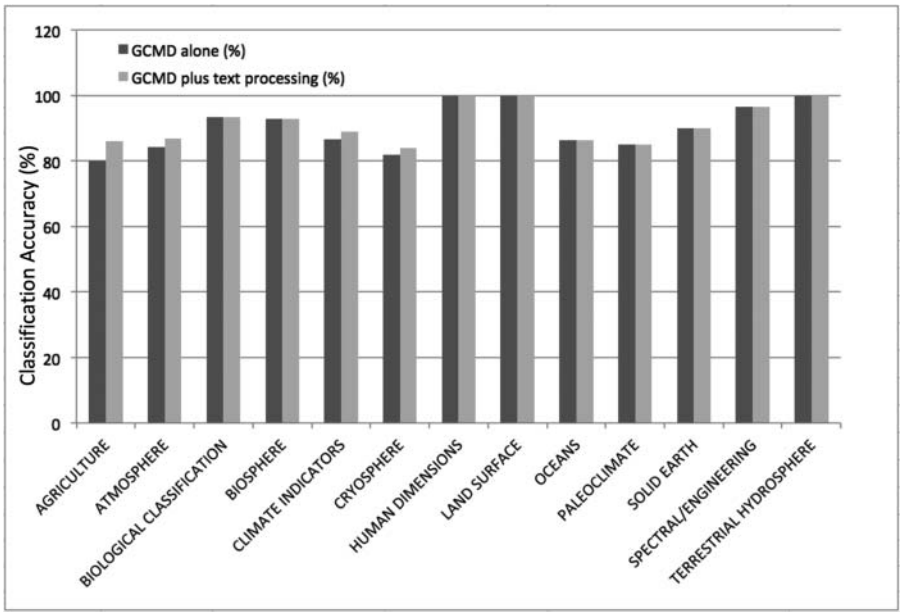

Figure 11. Classification accuracy comparison using Global Change Master Directory taxonomy alone and the combined taxonomy and text processing approach. GCMD = Global Change Master Directory.

the thirteen categories. Note that three categories (Human Dimensions, Land Surface, and Terrestrial Hydrosphere) received 100 percent classification accuracy for both approaches. The hybrid approach achieves better performance because the phrases extracted from N-gram analysis tend to contain more scientific meanings and therefore help to eliminate the ambiguity caused by matching a single keyword.

In summary, the set of experiments shows the performance boost as well as the high classification accuracy using the combined ontology and semantic matching approach to support thematic classification of records. This technique can better achieve the goal of helping researchers find the most data they need and at the same time increase the accessibility of data resources.

Performance of Space–Time Filter

To evaluate the performance of the space–time filter, thirty data sets providing data for "Greenland," "Australia," "California" (referring to California in the United States), and "Alaska" (referring to Alaska in the United States) were randomly selected and a comparison on recall and precision was conducted. Recall measures the portion of relevant data sets that are able to retrieve by the proposed space–time filter. Precision assesses the ratio between correctly retrieved data sets and all data sets returned by the proposed filter. Their equations are shown here. Table 1 shows the evaluation results.

Table 1. Comparison of precision and recall for the sample spatial queries using place-name as keyword

	Precision (%)		Recall (%)	
Query	Original	New	Original	New
Alaska	74.3	100.0	86.7	96.7
California	75.0	93.3	90.0	96.7
Greenland	78.7	100.0	86.7	100.0
Australia	74.3	90.0	86.7	100.0

$$Precison = \frac{\#\ of\ relevant\ dataset\ in\ the\ result\ set}{\#\ of\ result\ set}$$

$$Recall = \frac{\#\ of\ relevant\ dataset\ in\ the\ result\ set}{Total\ \#\ of\ relevant\ datasets}$$

As seen, after applying the new approach, both the recall and precision rates are improved substantially. The nonperfect recall rates for Alaska and California are due to lack of place-name mentioning in the metadata of the sampled data set, whose bounding box information is also missing. The precision rates show that the new approach more accurately defines the spatial coverage of the data set.

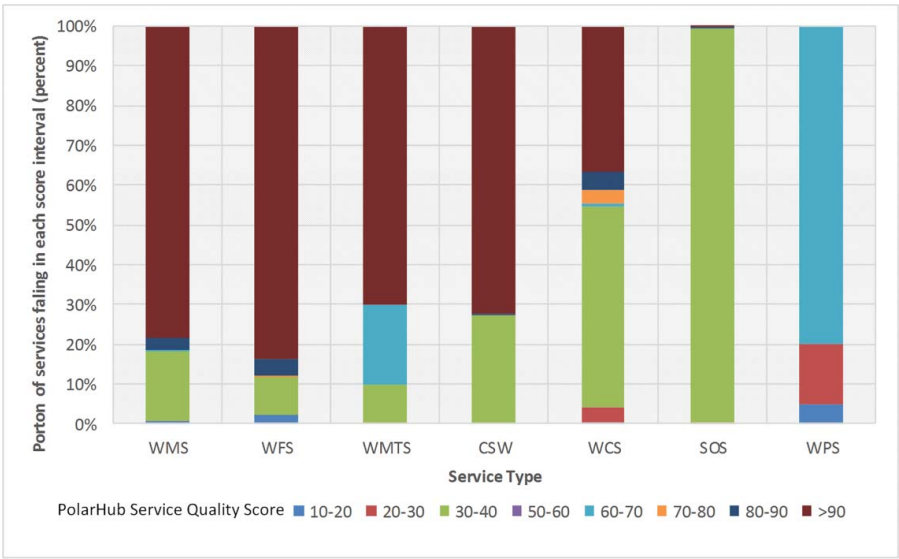

Figure 12. Portion of services falling in different score interval. *Note:* WMS = Web Map Service; WFS = Web Feature Service; WMTS = Web Map Tile Service; CSW = Catalogue Service for the Web; WCS = Web Coverage Service; SOS = Sensor Observation Service; WPS = Web Processing Service. (Color figure available online.)

Service Quality Evaluation Result

The quality of the data services found in PolarHub was further evaluated. The higher the score is, the better quality a service has. Figure 12 illustrates the portion of services falling in each score interval.

Results show that for all live WMS collected, a majority (~78 percent) received a score higher than ninety. Over 80 percent of the live WFSs received a score higher than ninety. About 70 percent of WMTS and CSW received a score higher than ninety. Only 37 percent of WCS received high scores, however, and very few of SOS received a score higher than eighty. Most SOS and WCS received a score between thirty and forty.

This discrepancy might be due to WM(T)S and WFS, the two most popular geospatial Web services, being widely supported by numerous high-performance Web platforms such as GeoServer, ESRI ArcGIS server, and so on, for service publishing. CSW also presents good stability and robustness across the data providers, as a popular service-oriented catalog solution. In contrast, the platforms for publishing SOS and WCS are relatively fewer. Therefore, WMS, WFS, and CSW servers tend to be more stable than SOS and WCS servers. In addition, the amount of data being transferred through SOS and WCS is usually higher than that transferred by WMS and WFS, because of the real-time characteristics of sensor observation data as well as the huge amount of coverage data. The amount of big data slowed down the server response time, causing a lower score in quality evaluation.

PolarHub Graphical User Interface

The PolarHub globe displays the clustered locations of all the Web services that have been crawled and saved in the back-end data clearinghouse. By clicking a bubble on the 3D globe, the detailed service data and metadata can be viewed. Figure 13 demonstrates the PolarHub GUI crawling system.

Besides providing the crawling information on the GUI, a search interface has also been developed. This allows an authorized user to discover additional data from the PolarHub data repository by providing keywords that generate a new crawling task initiated as a daemon program (i.e., a program that runs in the background). The end user can also refine the search results by selecting data from different organizations, quality, and other criteria. The results are ranked by keyword matching in the metadata fields of keywords, title, and abstract. Furthermore, PolarHub integrates methods for identifying the theme (shown as a category on the GUI), location, and service quality score via a new window when a specific service is clicked on the virtual globe.

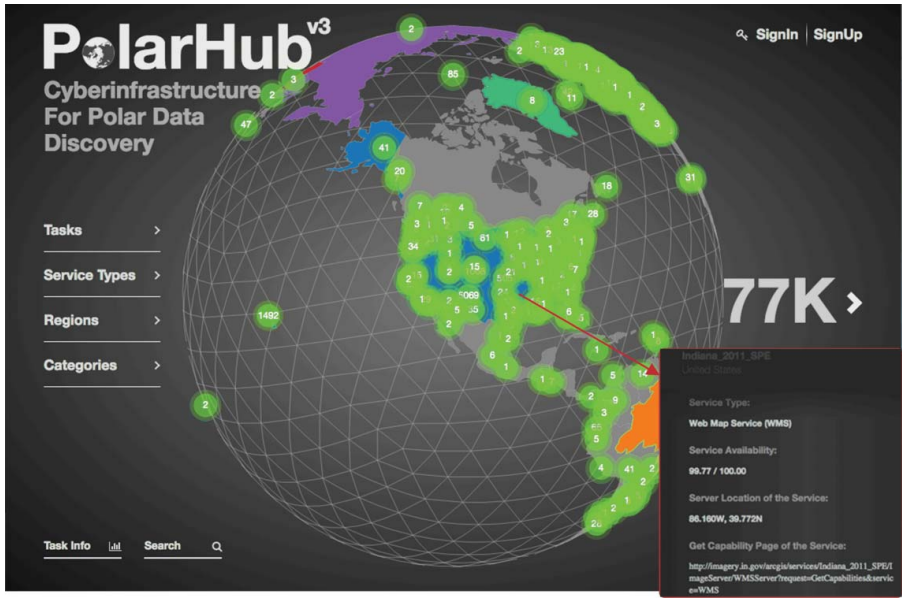

Figure 13. PolarHub graphical user interface (http://cici.lab.asu.edu/polarhub3). (Color figure available online.)

Conclusion and Discussion

This article introduces a novel solution to resolve the accessibility challenge that commonly exists in a number of science domains, especially the geospatial sciences in which data about locations are collected and analyzed. Its unique contributions include (1) a formal and comprehensive definition of a fundamental research topic in cyberinfrastructure research—the data access problem and its five unique and indispensable criteria for evaluating data accessibility; (2) Web-scale data crawling that enables the effective discovery of distributed geospatial data as well as data shared as services, thereby bridging the gap between data providers and data users; (3) the introduction of semantic analysis and spatial filtering techniques based on domain ontologies and gazetteers that provide a thorough analysis and a more accurate measure of the theme and geographical coverage of the spatial data sets (these techniques enable researchers to easily identify data sets that satisfy various scientific analysis needs); (4) the adoption and enhancement of an FGDC service quality checker that provides on-the-fly reliability evaluation of the geospatial data services; and (5) the seamless integration of the proposed methods into an operational crawling platform, the PolarHub, which continuously inspects the Web footprint of existing geospatial data.

Although the topic of Web crawling is not new, expanding this technique to make it suitable for geospatial data discovery is of great importance to advance the emerging spatial data science. Integrating it with comprehensive semantic and spatial analysis in an operational system environment provides a scientific analysis tool that allows researchers to better understand the changing distribution patterns of the ever-increasing geospatial data on the Web. In comparison to other spatial data infrastructure solutions, the PolarHub tool enables much more Web coverage of geospatial data. Furthermore, its data storage is continuing to increase with the data fully analyzed to identify staleness and update data content. To this end, PolarHub serves as an excellent testbed for various science applications as well as assessing geospatial interoperability and trend analysis of its adoption in the open GIScience community.

In the future, improvements will be further made to the techniques and methods. Several areas are already under investigation. On the computation side, the back-end computing paradigm will be extended from a multithreading model on a single compute node to a parallel model that uses the national cutting-edge, high-performance computing facility Resource Open Geo-Spatial Education and Research (ROGER) to achieve high efficiency in data crawling. Meanwhile, a series of experiments were being conducted to evaluate the performance in terms of system latency. The main source of latency comes from the initialization time that a server takes to allocate computing resources for a crawling task and the waiting time interval (10 seconds) to ping Web pages from the same remote server due to politeness policy settings. The statistics on thirty-four experiments (thirty-four different crawling tasks using a sixteen-thread compute model) show that on average the system latency (from starting a task to the first Web service to be found) is only 19.87 seconds. This result reflects good performance in terms of quickness and low latency in identifying geospatial data services.

In addition, the thematic classification framework is being enhanced to further improve its classification accuracy. From the experimental results in Figures 10 and 11, it can be seen that although the classification accuracy is relatively high and the hybrid approach works better than the ontology approach alone, there is room for improvement. Efforts are being made to develop a new automated approach to remove the ambiguity in the scientific taxonomy and the efficiency in text processing. Strategies are also being investigated to improve the search functionality and provide intelligent search capabilities of data within PolarHub on a finer granularity. The goal is to allow searching individual observation sites, such as an SOS, rather than searching an entire service. PolarHub's search scope will be extended to cover a broader range of geospatial data types, such as Shapefiles, CSV, and GeoJSON.

Acknowledgments

Assistance received from Sizhe Wang on data processing is greatly appreciated. The author would also like to thank the editor and anonymous reviewers for their valuable and constructive comments.

Funding

This article draws on work supported in part by the following awards: PLR-1349259; BCS-1455349; and PLR-1504432 from the National Science Foundation

and another award from the Open Geospatial Consortium.

References

Agarwal, P. 2005. Ontological considerations in GIScience. *International Journal of Geographical Information Science* 19 (5):501–36.

Allard, S. 2012. Data one: Facilitating eScience through collaboration. *Journal of eScience Librarianship* 1 (1): e1004. http://dx.doi.org/10.7191/jeslib.2012.1004

Ames, D. P., J. S. Horsburgh, Y. Cao, J. Kadlec, T. Whiteaker, and D. Valentine. 2012. Hydro desktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling and Software* 37:146–56.

Atkins, D. E., K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein, D. G. Messerschmitt, and M. H. Wright. 2003. Final report of the NSF blue ribbon advisory panel on cyberinfrastructure: Revolutionizing science and engineering through cyberinfrastructure. Accessed April 14, 2017. https://www.nsf.gov/cise/sci/reports/atkins.pdf

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. *Dbpedia: A nucleus for a web of open data*. New York: Springer.

Avery, S. 2000. *NSF geosciences beyond 2000: Understanding and predicting Earth's environment and habitability*. Arlington, VA: Directorate for Geosciences, National Science Foundation. Accessed April 20, 2017. https://www.nsf.gov/geo/adgeo/geo2000.jsp

Bai, Y., L. Di, D. D. Nebert, A. Chen, Y. Wei, X. Cheng, and H. Wang. 2012. GEOSS component and service registry: Design, implementation and lessons learned. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (6):1678–86.

Bell, G., T. Hey, and A. Szalay. 2009. Beyond the data deluge. *Science* 323 (5919):1297–98.

Bergman, M. K. 2001. White paper: The deep web: Surfacing hidden value. *Journal of Electronic Publishing* 7 (1). Accessed October 2, 2017. http://dx.doi.org/10.3998/3336451.0007.104

Bespalov, D., B. Bai, Y. Qi, and A. Shokoufandeh, eds. 2011. Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ed. B. Berendt, A. de Vries, W. Fan, C. Macdonald, I. Ounis, and I. Ruthven, 375–82. New York: ACM.

Biettron, L., F. Pallu, and S. Tricot. 2006. U.S. Patent No. 7,003,519, filed September 22, 2000, and issued February 21, 2006.

Bishr, Y. 1998. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science* 12 (4):299–314.

Bukhres, O., Z. B. Miled, E. Lynch, L. Olsen, and Z. Tari. 2000. Effective standards for metadata in the GCMD data access system. In *Distributed Objects and Applications, 2000 proceedings*, ed. P. Drew, 155–61. Washington, DC: IEEE.

Christian, E. 2005. Planning for the global earth observation system of systems (GEOSS). *Space Policy* 21 (2):105–9.

Clark, M., and S. Watt, eds. 2007. *Classifying XML documents by using genre features: DEXA'07 18th International Workshop on Database and Expert Systems Applications*. Berlin Heidelberg: Springer-Verlag.

de La Beaujardiere, J. 2006. OpenGIS® Web map server implementation specification. OGC 06-042, Open Geospatial Consortium. Accessed October 2, 2017. http://portal.opengeospatial.org/files/?artifact_id=14416

ESRI. 2011. ESRI ArcInfo grid. Accessed October 2, 2017. https://www.loc.gov/preservation/digital/formats/fdd/fdd000281.shtml.

Fonseca, F., M. Egenhofer, C. Davis, and G. Câmara. 2002. Semantic granularity in ontology-driven geographic information systems. *Annals of Mathematics and Artificial Intelligence* 36 (1–2):121–51.

Foster, I. 2005. Service-oriented science. *Science* 308 (5723):814–17.

Fox, P., and J. Hendler. 2014. The science of data science. *Big Data* 2 (2):68–70.

Ganjisaffar, Y. 2012. Crawler4j: Open source Web crawler for Java. Accessed October 2, 2017. https://github.com/yasserg/crawler4j

Gao, S., L. Li, W. Li, K. Janowicz, and Y. Zhang. 2014. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems* 61 (B):172–86.

Global Change Master Directory (GCMD). 2015. *GCMD keywords, version 8.1*. Greenbelt, MD: Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center National Aeronautics and Space Administration. http://gcmd.nasa.gov/learn/keywords.html

Gold, A. K. 2007. Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *Office of the Dean (Library)* 13 (9/10):17

Goodchild, M. F. 2007. Citizens as sensors: The world of volunteered geography. *Geo Journal* 69 (4):211–21.

———. 2013. Geocoding and geosampling. *Spatial Statistics and Models* 40:33–53.

Goodchild, M. F., and D. G. Janelle. 2010. Toward critical spatial thinking in the social sciences and humanities. *GeoJournal* 75 (1):3–13.

Halevy, A. 2005. Why your data won't mix. *Queue* 3 (8):50–58.

Hey, A. J., S. Tansley, and K. M. Tolle. 2009. *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Huang, C. Y., and H. Chang. 2016. GeoWeb crawler: An extensible and scalable Web crawling framework for discovering geospatial Web resources. *ISPRS International Journal of Geo-Information* 5 (8):136.

INSPIRE E. Directive. 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*.

Janowicz, K., and P. Hitzler. 2013. Thoughts on the complex relation between linked data, semantic

annotations, and ontologies. In *Proceedings of the sixth international workshop on exploiting semantic annotations in information retrieval*, ed. P. N. Bennett, E. Gabrilovich, J. Kamps, J. Karlgren, 41–44. New York: ACM.

Kwan, M. P. 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers* 106 (2):274–82.

Lakhani, K. R., R. D. Austin, and Y. Yi. 2010. *Data.gov*. Cambridge, MA: Harvard Business School.

Li, W., M. F. Goodchild, and R. Raskin. 2014. Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth* 7 (1):17–37.

Li, W., S. Wang, and V. Bhatia. 2016. Polar Hub: A large-scale web crawling engine for OGC service discovery in cyberinfrastructure. *Computers, Environment and Urban Systems* 59:195–207.

Li, W., C. Yang, D. Nebert, R. Raskin, P. Houser, H. Wu, and Z. Li. 2011. Semantic-based web service discovery and chaining for building an Arctic spatial data infrastructure. *Computers & Geosciences* 37 (11):1752–62.

Li, W., C. Yang, and R. Raskin. 2008. A semantic enhanced search for spatial web portals. In *AAAI Spring Symposium: Semantic scientific knowledge integration*, ed. D. L. McGuinness and P. Fox, 47–50. Palo Alto, CA: AAAI.

———. 2009. A semantic-enabled meta-catalogue for intelligent geospatial information discovery. In *17th International Conference on Geoinformatics*, ed. L. Di and A. Chen, 1–5. Fairfax, VA: IEEE.

Li, W., C. Yang, and C. Yang. 2010. An active crawler for discovering geospatial web services and their distribution pattern—A case study of OGC Web Map Service. *International Journal of Geographical Information Science* 24 (8):1127–47.

Lü, X., L. Zhang, and J. Hu. 2004. Statistical substring reduction in linear time. In *Natural language processing—IJCNLP 2004*, 320–27. Berlin, Germany: Springer.

Lutz, M., J. Sprado, E. Klien, C. Schubert, and I. Christ. 2009. Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences* 35 (4):739–52.

MacEachren, A. M. 1991. The role of maps in spatial knowledge acquisition. *The Cartographic Journal* 28 (2):152–62.

Maguire, D. J., and P. A. Longley. 2005. The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment, and Urban Systems* 29 (1): 3–14.

Mani, A., and A. Nagarajan. 2005. Understanding quality of service for Web services. Accessed October 2, 2017. https://www.ibm.com/developerworks/library/ws-quality/index.html

Masó, J., K. Pomakis, and N. Julià. 2010. OGC Web Map Tile Service (WMTS), Implementation standard, OGC 07-057r7. Accessed October 2, 2017. http://portal.opengeospatial.org/files/?artifact_id=35326

Michener, W., D. Vieglais, T. Vision, J. Kunze, P. Cruse, and G. Janée. 2011. Dataone: Data observation network for earth-preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine* 17 (1):3. doi:10.1045/january2011-michener

Miled, Z. B., S. Sikkupparbathyam, O. Bukhres, K. Nagendra, E. Lynch, M. Areal, L. Olsen, C. Gokey, D. Kendig, T. Northcutt, et al. 2001. Global change master directory: Object-oriented active asynchronous transaction management in a federated environment using data agents. In *Proceedings of the 2001 ACM Symposium on Applied Computing*, ed. G. B. Lamont, 207–14. New York: ACM.

Miller, H. J. 2010. The data avalanche is here: Shouldn't we be digging? *Journal of Regional Science* 50 (1):181–201.

Miller, H. J., and M. F. Goodchild. 2015. Data-driven geography. *GeoJournal* 80 (4):449–61.

Na, A., and M. Priest. 2007. Sensor observation service. Implementation Standard. OGC 06-009r6. Accessed October 2, 2017. http://portal.opengeospatial.org/files/?artifact_id=26667

Nagao, M., and S. Mori. 1994. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In *Proceedings of the 15th conference on Computational Linguistics*, ed. M. Nagao and Y. Wilks, Vol. 2, 611–15. Stroudsburg, PA: Association for Computational Linguistics.

New, M., M. Hulme, and P. Jones. 2000. Representing twentieth-century space–time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. *Journal of Climate* 13 (13):2217–38.

Ramamurthy, M. K. 2006. A new generation of cyberinfrastructure and data services for earth system science education and research. *Advances in Geosciences* 8 (8):69–78.

Raskin, R. G., and M. J. Pan. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences* 31 (9):1119–25.

Reid, W. V., D. Chen, L. Goldfarb, H. Hackmann, Y. T. Lee, K. Mokhele, and A. Whyte. 2010. Earth system science for global sustainability: Grand challenges. *Science* 330 (6006):916–17.

Ritter, N., and M. Ruth. 2000. GeoTIFF format specification. GeoTIFF revision 1.0, Spec v1.8.2, 28 December. Accessed October 2, 2017. https://www.loc.gov/preservation/digital/formats/fdd/fdd000279.shtml

Rosenfield, G. H., and K. Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing* 52 (2):223–27.

Scholl, M., and A. Voisard. 1990. Thematic map modeling. In *Design and implementation of large spatial databases*, ed. A. P. Buchmann, O. Günther, T. R. Smith, Y. F. Wang, 167–90. New York: Springer.

Sicilia, M. A. 2006. Metadata, semantics, and ontology: Providing meaning to information resources. *International Journal of Metadata, Semantics and Ontologies* 1 (1):83–86.

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data sharing by scientists: Practices and perceptions. *PloS ONE* 6 (6):e21101.

Vatant, B., and M. Wick 2006. Geonames ontology. Accessed October 2, 2017. http://www.geonames.org/ontology

Vatsavai, R. R., and B. Bhaduri. 2011. A hybrid classification scheme for mining multisource geospatial data. *GeoInformatica* 15 (1):29–47.

Ventrone, V. 1991. Semantic heterogeneity as a result of domain evolution. *ACM SIGMOD Record* 20 (4):16–20.

Vretanos, P. A. 2005. Web feature service implementation specification. *Open Geospatial Consortium Specification* 1325 04-094. Accessed October 2, 2017. http://portal.opengeospatial.org/files/?artifact_id=8339

Wang, Y., T. Wang, X. Ye, J. Zhu, and J. Lee. 2015. Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm. *Sustainability* 8 (1):25. doi:10.3390/su8010025

Whitehouse. 2012. Whitehouse Big Data initiatives. Accessed September 30, 2017. https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200.

Whiteside, A. 2007. OGC Web services common specification. OGC document 06-121r9. Accessed October 2, 2017. http://portal.opengeospatial.org/files/?artifact_id=38867

Whiteside, A., and J. D. Evans. 2008. Web Coverage Service (WCS) implementation standard. OGC document 08-059r4. Accessed October 2, 2017. https://portal.opengeospatial.org/files/08-059r4

Widener, M. J., and W. Li. 2014. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography* 54:189–97.

Worboys, M. F., and S. M. Deen. 1991. Semantic heterogeneity in distributed geographic databases. *ACM SIGMOD Record* 20 (4):30–34.

Xia, J., C. Yang, K. Liu, Z. Li, M. Sun, and M. Yu. 2015. Forming a global monitoring mechanism and a spatiotemporal performance model for geospatial services. *International Journal of Geographical Information Science* 29 (3):375–96.

Yin, J., J. T. Overpeck, S. M. Griffies, A. Hu, J. L. Russell, and R. J. Stouffer. 2011. Different magnitudes of projected subsurface ocean warming around Greenland and Antarctica. *Nature Geoscience* 4 (8):524–28.

WENWEN LI is an Associate Professor in the School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287–5302. E-mail: Wenwen@asu.edu. Her research interests include methodology development in cyberinfrastructure, spatial data science, geospatial semantics, deep learning and their applications in polar environmental change, terrain analysis, sustainability, and urban heat island research.