[DM-07-060] Spatial Data Infrastructures

Abstract

Spatial data infrastructure (SDI) is the infrastructure that facilitates the discovery, access, management, distribution, reuse, and preservation of digital geospatial resources. These resources may include maps, data, geospatial services, and tools. As cyberinfrastructures, SDIs are similar to other infrastructures, such as water supplies and transportation networks, since they play fundamental roles in many aspects of the society. These roles have become even more significant in today's big data age, when a large volume of geospatial data and Web services are available. From a technological perspective, SDIs mainly consist of data, hardware, and software. However, a truly functional SDI also needs the efforts of people, supports from organizations, government policies, data and software standards, and many others. In this chapter, we will present the concepts and values of SDIs, as well as a brief history of SDI development in the U.S. We will also discuss the components of a typical SDI, and will specifically focus on three key components: geoportals, metadata, and search functions. Examples of the existing SDI implementations will also be discussed.

Keywords: data standards, geoportal, geospatial interoperability, institutional and inter-institutional aspects, metadata, Web service

Author & citation

Hu, Y. and Li, W. (2017). Spatial Data Infrastructures. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2017 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2017.2.1

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Spatial data infrastructures. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions
2. Spatial Data Infrastructure and Its Values
3. Key Components of an SDI
4. Examples of SDIs
5. Current and Possible Future Possible Developments of SDIs

1. Definitions

spatial data infrastructure: The technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data, services, and other digital resources.

geoportal: A gateway website through which people can search, discover, access, and visualize the geospatial resources within a SDI.

metadata: Documentation about who, when, how, what, why, and many other facets of the data and the data production process. Metadata can be used for describing not only data, but also tools, services, and other geospatial resources.

data standard: A commonly agreed specification on how data should be recorded and described.

geospatial interoperability: The ability of different geographic information systems to share, exchange, and operate (heterogenous) geospatial data and functions.

web service: A Web application that provides standardized application programming interfaces to allow remote access to data and functions over the Internet.

2. Spatial Data Infrastructure and its Values

The emergence of spatial data infrastructures (SDIs) is closely associated with the efforts of collecting and producing geospatial data, as well as the advancement of surveying and computer technologies. In the past decades, a large amount of geospatial data, such as remote sensing images and GPS locations, have been collected by government agencies such as the U.S. Geological Survey (USGS) and the National Oceanic and Atmospheric Administration (NOAA). Meanwhile, the fast development of geographic information systems facilitates the derivation of various data products from the collected data, such as topographic maps, land cover data, transportation networks, and hydrographic features. As location-based services are becoming increasingly popular, vast amounts of volunteered geographic information (VGI) (Goodchild, 2007) has also been contributed by the general public through smart mobile devices and social media platforms. In addition, the componentization of GIS brings geospatial services that provide data processing and spatial analysis functions in the general Web environment. The large number of geospatial data, services, maps, and others, however, do not ease the use of these geospatial resources. On one hand, it is challenging to find and access these digital resources which are widely distributed at different government agencies and websites (Li, Wang, & Bhatia, 2016). On the other hand, a lot of data redundancies exist, and money and human resources were wasted in duplicated data collection and maintenance efforts (Rajabifard & Williamson, 2001; Maguire & Longley, 2005).

These problems were recognized by governments of the countries around the world, and many spatial data infrastructures were constructed since 1990s (Masser, 1999). In the U.S., a national spatial data infrastructure (NSDI) initiative was started in 1993 to provide standardized access to geographic information resources (National Research Council, 1993). An official definition of NSDI, according to the Executive Order 12906, is “the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data.” The Federal Geographic Data Committee (FGDC) is charged with coordinating the efforts to develop the NSDI in the U.S. The naming also indicates that SDI is recognized as an infrastructure similar to other types of infrastructures, such as electricity grids and water supplies, and that it plays fundamental roles in the socioeconomic and environmental developments of a country.

Three parallel fronts were developed in the NSDI program: 1) a set of data standards for formalizing data and metadata; 2) a clearinghouse network providing data storage and online access; and 3) a set of framework data for the entire country, such as administrative boundaries (Longley et al., 2001; Maguire & Longley, 2005).

Spatial data infrastructure presents a solution to the problems of resource discovery and data redundancy. It provides a unified platform where people can go and search geospatial data, maps, services, and other digital resources. As multiple government agencies are sharing their data on one platform, SDI reduces data redundancy and the extra efforts in collecting duplicated geospatial data. From a cost/benefit perspective, SDI allows geospatial data to be collected once and reused multiple times in different applications. More generally, SDI can be considered as an important element in the e-government (Georgiadou, Rodriguez-Pabón, & Lance, 2006)and open-government movement to increase the transparency of governmental activities and to enhance public participation. Better access to geospatial data also stimulates the growth of new businesses which may not be possible otherwise (Ralston, 2004).

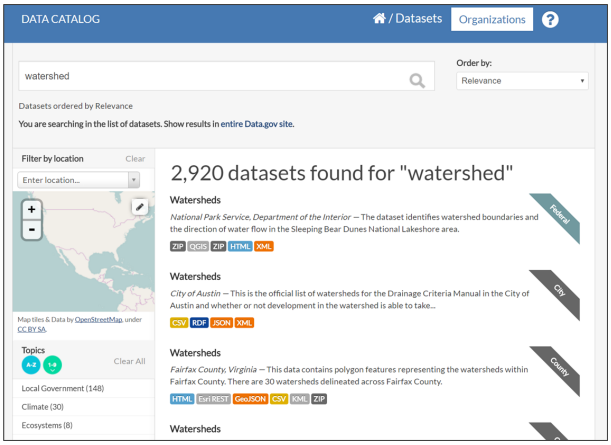
3. Key Components of an SDI

An SDI consists of many components. In addition to the digital geospatial resources, a SDI also needs hardware, software, people, organizations, standards, policies, and many others to function properly. Constructing a SDI also needs effective communications between communities, and negotiations among organizations and even countries to reach agreements. While a SDI has many components, this chapter will particularly focus on geoportals, metadata, and search functions, which are three key components of a typical SDI.

3.1 Geoportals

Geoportals are Web gateways that provide one-stop access to geospatial resources (Tait, 2005). Geoportals are probably the most visible part of SDIs, since they are the main interfaces through which people can search and find geospatial resources. Figure 1 shows the user interface of Data.gov, a geoportal for the U.S. government open data. Figure 1(a) is the initial interface of the geoportal which contains a search bar that allows users to type in some keywords and find relevant resources. Figure 1(b) shows a list of the results when the keyword “watershed” is searched.





(a)

(b)

Figure 1. Two screenshots of the U.S. government open data geoportal Data.gov.

Geoportals are typically developed using Web-based technologies and off-the-shelf GIS software packages. A database management system (DBMS) is used to store and manage the metadata of the geospatial resources contained in the SDI. A Web interface, which often contains a map, enables end users to interact with the system and to conduct searches (Figure 2). When a search is performed, a HTTP (Hypertext Transmission Protocol) request will be sent to the Web server which hosts the geoportal. After querying the metadata stored in the database, the geoportal will then send back the result to the client through a HTTP response. Geoportals are typically designed to be used by both GIS professionals and the general public.

One important function of geoportals is helping users discover the existing geospatial resources. This resource discovery process often follows the publish-find-bind pattern (Rose, 2004; Maguire & Longley, 2005), in which: 1) providers publish the metadata of their data and services to a geoportal; 2) users perform a search on the geoportal and potentially find the data; 3) users consume the data and services from the providers. Figure 2 illustrates these three steps.

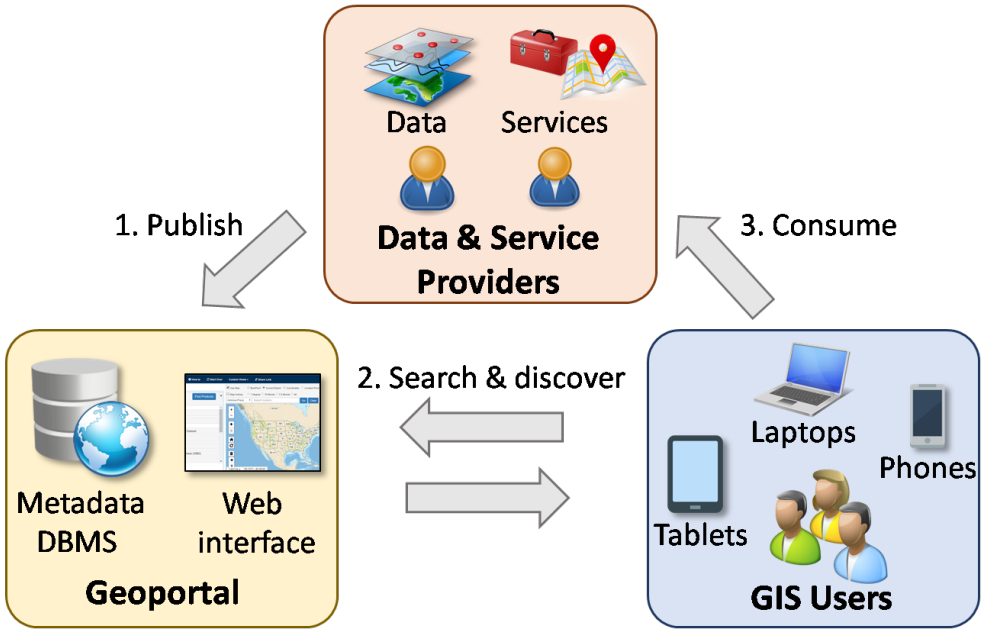


Figure 2. Some key components of an SDI and the publish-find-bind pattern.

The resource discovery process heavily relies on the quality of the metadata and the performance of the search function. In the following, we will discuss these two components.

3.2 Metadata

Metadata provide documentations on the content and the production process of geospatial

resources. Metadata are often called the data about data, and include information such as titles, descriptions, data categories, the locations and time of the data collection, the data collectors, the used coordinate systems and map projections, and the data cleaning and processing procedures. Metadata can also be used for describing geospatial services by providing information about the data and functions offered by the services, the input and output, the developers, the development time, and others. In short, metadata are about all aspects of digital geospatial resources. Figure 3 shows the metadata fragment of a geospatial service provided by the NOAA, which describes the capabilities (e.g., getting a map based on the served geospatial data) and the data layers provided by this service.



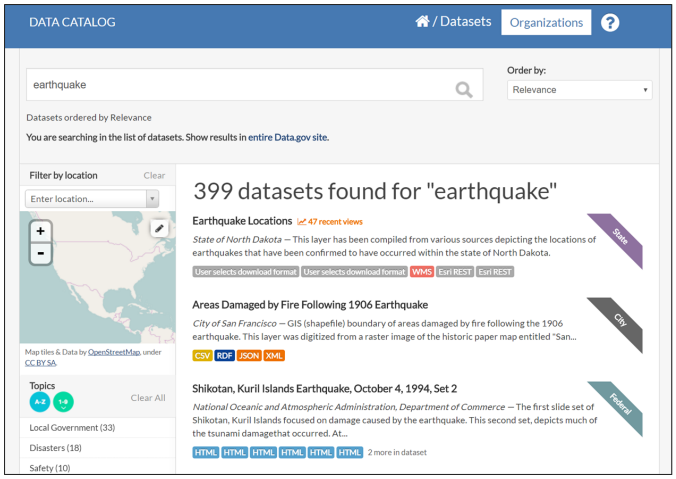
Figure 3. A metadata fragment (in the format of XML) of a geospatial service provided by the NOAA.

Metadata are fundamentally important for SDIs. When data and services leave their original data production context and are integrated into a SDI, metadata provide the primary information based on which GIS users can understand and use digital geospatial resources. Without metadata or with only poorly constructed metadata, it is very difficult, if not impossible, for data and services to be reused. The quality of metadata also determines the result of resource discovery. Many geoportals rank the relevance of geospatial resources to user queries based on the information contained in their metadata. Complete and accurate metadata allow geoportals to find and rank geospatial resources based on locations, time, thematic attributes, data types, published years, data collectors, and many other conditions explicitly or implicitly specified in the user queries. To ensure the quality of metadata, standards are established to define the necessary elements that should be included in metadata. In the U.S., the FGDC is responsible for coordinating the development of metadata standards, and its Content Standard for Digital Geospatial Metadata (CSDGM) has been used by many U.S. government agencies to formalize metadata. Since 2010, FGDC has endorsed a series of international metadata standards (e.g., the ISO 191** standards) to promote Global Spatial Data Infrastructure (GSDI).

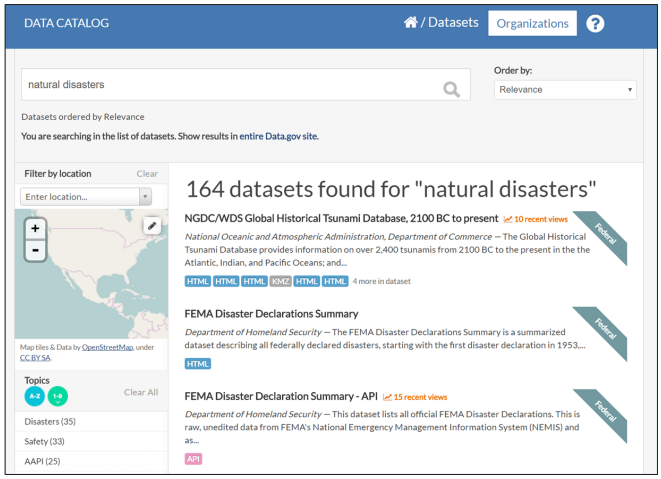
3.3 Search Function

Search functions are the major means through which users discover geospatial resources in a SDI. Without an effective search function, relevant geospatial resources in an SDI can hardly be found by the users. Two types of search functions are often adopted in geoportals: text-based search and map-based search. Text-based search is similar to Web search engines, in which a user types in some keywords and receives the results based on the matched text. Map-based search allows users to find geospatial resources by interacting with a map, and a user can pan, zoom in and out, and draw polygons to specify their areas of interest. Using either of these two methods alone has some limitations. The text-based search enables general users, especially the users who are unfamiliar with a GIS, to find geospatial resources in a way similar to how they would use a general search engine, such as Google. However, it can be challenging to identify the suitable keywords, or the keywords may not accurately describe the geographic areas that the user is interested in. The map-based search, on the other hand, provides convenience for the users who are already familiar with a map interface, and allows users to specify geographic locations in a more accurate manner (e.g., by drawing polygons). However, not everyone feels comfortable of using a map-based interface. Geoportals often offer both search functions to complement the two and accommodate the needs of different users.

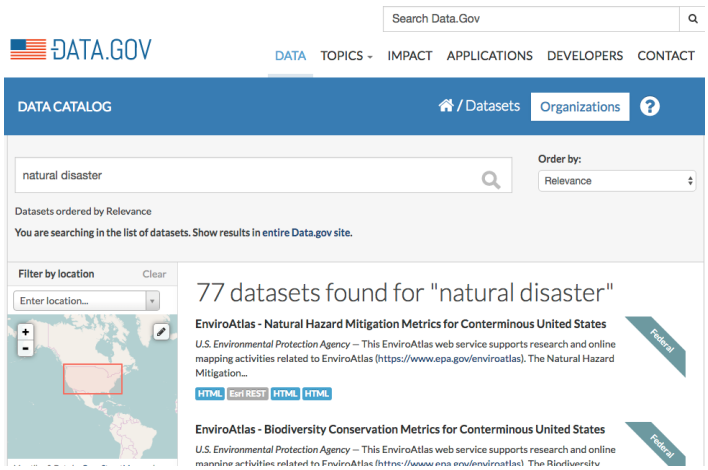
The search functions in SDIs are being improved by researchers. One of these improvements lies in text-based search: there is a transition from keyword-based search to semantic search. Keyword-based search examines the matching between the keywords input by the users and the textual descriptions of the geospatial resources. Thus, if a user types in “road”, the search function will not be able to find the resources labeled as “street”. Semantic search aims to match digital geospatial resources to user queries based on the meaning (semantics) of the queries, and therefore can identify relevant data and services even though they are not labeled with the exactly same words. Early attempts of semantic search include the use of ontology-based query expansion (Lutz & Klien, 2006), rule-based semantic reasoning (Li, Zhou, & Wu, 2016), and faceted search, e.g., Apache Solr (Hostetter, 2006). Despite the existing research efforts, additional work is still necessary to integrate semantic search into current SDIs. Figure 4 shows examples of searching “earthquake” (Figure 4(a)) and “natural disasters” (Figure 4(b)) respectively on Data.gov. While earthquake is generally considered as a type of natural disaster, the search of “earthquake”, however, returned more matching records than the other search, suggesting that the search function behind is more likely to be based on a simple keyword matching. In addition to text-based search, map-based search is also supported by Data.gov. Figure 4(c) shows a spatial filtering effect by narrowing down the region of interest to the contiguous United States using the map interface.



(a)



(b)



(c)

Figure 4. Search examples based on Data.gov.

4. Examples of SDIs

There are many spatial data infrastructures established at different geographic levels. At the global level, there is Global Earth Observation System of Systems which combines the efforts from more than 70 countries to share environmental data. At the continental level, there is Infrastructure for Spatial Information in the European Community (INSPIRE) which enables the sharing of geospatial information among public organizations across Europe. At the national level, there is Data.gov which provides access to a large number of United States government open datasets. The National Map (TNM) is a USGS SDI project which supports easy access and downloading of topographic information about elevation, geographic names, hydrology, boundaries, transportation, and so forth. There are also national SDIs in Australia, China, Japan, Malaysia, Netherlands, Portugal, and other countries. At the state level, there is Tennessee GIS portal which offers the geospatial datasets, services, and Web applications related to the State of Tennessee. At the city level, there is New York City Open Data portal . There are also spatial data infrastructures dedicated to specific domains, such as disaster response, public health, and climate

change.

5. Current and Possible Future Developments of SDIs

Spatial data infrastructures heavily rely on computer and information technologies, and are continuously evolving with the technological advancements. Techniques commonly used in today's SDIs, such as Asynchronous JavaScript and XML (AJAX) which enables asynchronous processing to speed up search performance as well as user experience, are not available for the first generation of SDIs in 1990s. Similarly, we may see the emergence of new technologies that can improve SDIs in various aspects, and some of these technologies are already being tested in research labs. For example, graph databases, such as Neo4j, may be employed to augment the existing relational databases that have been widely used in SDIs to store metadata. Semantic Web (or the third generation of the Web) and Linked Data technologies may be leveraged to turn SDIs into local Semantic Webs (Athanasis et al., 2009). The concept of knowledge graph (Singhal, 2012) may be applied and integrated to SDIs to construct knowledge bases that link the geospatial entities and concepts contained in the data. Machine learning and data mining methods, such as latent semantic analysis and labeled latent Dirichlet allocation may help automatically infer and generate high-quality metadata from the content of maps and services (Li, Bhatia, & Cao, 2015, Hu et al., 2015). While there can be lots of possible technological improvements, future SDIs also need the critical support from people, organizations, and governments. With these important components, a future SDI will contribute more to the development of the society

References

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

Georgiadou, Y., Rodriguez-Pabón, O., & Lance, K. T. (2006). Spatial data infrastructure (SDI) and e-governance: A quest for appropriate evaluation approaches. URISA-WASHINGTON DC, 18, 43.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221.

Hostetter, C. (2006). Faceted searching with apache solr. Presentation at ApacheCon US, 2006.

Hu, Y., Janowicz, K., Prasad, S., & Gao, S. (2015). Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online. Transactions in GIS, 19, 398-416.

Li, W., Bhatia, V., & Cao, K. (2015). Intelligent polar cyberinfrastructure: enabling semantic search in geospatial metadata catalogue to support polar data discovery. Earth Science Informatics, 8, 111-123.

Li, W., Wang, S., & Bhatia, V. (2016), PolarHub: A large-scale web crawling engine for OGC

service discovery in cyberinfrastructure. Computers, Environment and Urban Systems, 59, 195-207.

Li, W., Zhou, X., & Wu, S. (2016). An Integrated Software Framework to Support Semantic Modeling and Reasoning of Spatiotemporal Change of Geographical Objects: A Use Case of Land Use and Land Cover Change Study. ISPRS International Journal of Geo-Information, 5, 179.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2001). Geographic Information System and Science (1st Edition). England: John Wiley & Sons, Ltd.

Lutz, M., & Klien, E. (2006). Ontology-based retrieval of geographic information. International Journal of Geographical Information Science, 20, 233-260.

Maguire, D. J. & 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.

National Research Council (NRC). (1993). Toward a Coordinated Spatial Data Infrastructure for the Nation. Washington, D.C.: National Academy Press.

Rajabifard, A. & Williamson, I. P. (2001). Spatial data infrastructures: concept, SDI hierarchy and future directions. In GEOMATICS'80 Conference. Tehran, Iran.

Ralston, B. A. (2004). GIS and Public Data. Thomson/Delmar Learning.

Rose, L. (2004). Geospatial portal reference architecture: a community guide to implementing standards-based geospatial portals. OpenGIS Disscusion Paper, OGC, 04-039.

Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. The Keyword, the official Google blog.

Tait, M. G. (2005). Implementing geoportals: applications of distributed GIS. Computers, Environment and Urban Systems, 29, 33-47.