Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume II
Data Pro–I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

The Evolution of SDI Geospatial Data Clearinghouses

Maurie Caitlin Kelly
The Pennsylvania State University, USA

Bernd J. Haupt
The Pennsylvania State University, USA

Ryan E. Baxter
The Pennsylvania State University, USA

INTRODUCTION

Geospatial data and the technologies that drive them have altered the landscape of our understanding of the world around us. The data, software and services related to geospatial information have given us the opportunity to visualize existing phenomena, to understand connections, and to address problems from environmental management to emergency response. From the ever-present Google Earth images we are shown in our televised weather reports to the 3D flyovers of war zones on the news, geospatial information is everywhere. In the decade or so since U.S. President William Clinton set the stage by announcing the establishment of the National Spatial Data Infrastructure (NSDI), the concept of the geospatial data clearinghouse has shifted dramatically to fulfill the increasing need to streamline government processes, increase collaboration, and to meet the demands of data users and data developers (Clinton, 1994). The announcement of the NSDI gave birth to a Global Spatial Data Infrastructure (GSDI) movement that would be supported by a network of SDIs or geospatial data clearinghouses from local, state, and national levels.

From this point on, the evolution of the geospatial data clearinghouse has been rapid and punctuated with challenges to both the developer and the user. From the earliest incarnations of these now pervasive resources as simple FTP data transfer sites to the latest developments in Internet Map Services and real time data services, geospatial data clearinghouses have provided the backbone for the exponential growth of Geographic Information Systems (GIS). In this section, the authors will examine the background of the geospatial data clearinghouse movement, address the basic phases of clearinghouse development, and review the trends that have taken the world's clearinghouses from FTP to Internet Map Services and beyond.

THE SPATIAL DATA INFRASTRUCTURE MOVEMENT

No discussion of SDIs and geospatial data clearinghouses would be complete without a brief introduction to the history of the movement.

The growth of geospatial data clearinghouse movement can trace its origins to the spatial data infrastructure initiatives of the 1990s when spatial data sharing began in earnest. In the United States an effort to organize spatial data and develop standards for sharing data began as the NSDI. First envisioned in 1993, the concept of the coordinated data model set forth the ideas and goals of widespread sharing of data and resources (National Research Council, 1993). By 1995, the United States had developed a plan for data sharing and established a gateway by which participants could register their metadata holdings through a centralized source (FGDC95). Sharing data through this gateway required developing metadata to an accepted standard and utilized the Z39.50 protocol—both of which will be described in the next section.

The spatial data infrastructure concept as it has evolved has, at its core, the premise that sharing data eliminates redundancy, enhances opportunities for cooperative efforts, and facilitates collaboration. In addition, the SDI movement also has two additional advantages. First, it allows a more effective and efficient interaction with geospatial data and, second, it helps to stimulate the market for the geospatial industry

(Bernard, 2002). The general approach to developing an SDI is to first understand how and where geospatial data is created. Most SDIs or geospatial clearinghouses base their first level data collection efforts on framework data (FGDC95). Framework data is created by government agencies—local, state, federal, or regional for the purpose of conducting their business such as development and maintenance of roads, levying taxes, monitoring streams, or creating land use ordinances. These business practices translate themselves, in the geopspatial data world, into transportation network data, parcel or cadastral data, water quality data, aerial photographs, or interpreted satellite imagery. Other organizations can then build upon this framework data to create watershed assessments, economic development plans, or biodiversity and habitat maps. This pyramid of data sharing—from local to national—has been the cornerstone of the original concept of the SDI and considered a fundamental key to building an SDI (Rajabifard & Williamson, 2001).

The SDI movement now encompasses countries and regions all over the world and is now considered a global movement and potential global resource. Many countries maintain now clearinghouses participating in regional efforts. One effort along these lines is the GSDI (Nebert, 2004). The GSDI, which resulted from meetings held in 1995, is a non-profit organization working to further the goals of data sharing and to bring attention to the value of the SDI movement with a particular emphasis on developing nations (Stevens et al., 2004). Other projects including the Geographic Information Network in Europe (GINIE) project are working toward collaboration and cooperation in sharing geospatial data (Craglia, 2003). As of 2006, there were approximately 500 geospatial data clearinghouses throughout the world. The activities of the clearinghouses range from coordinating data acquisition and developing data standards to developing applications and services for public use with an average operating cost of approximately € 1,500,000 per year (approximately $ 1,875,000) (Crompvoets et al., 2006).

EVOLUTION OF SERVICES AND ACCESS IN THE GEOSPATIAL DATA CLEARINGHOUSE

There are several developmental phases that geospatial data clearinghouses engage in to become fully operational and integrated into a larger SDI, e.g., data acquisition and documentation, data access and retrieval capabilities, storage architecture development, and application development. These phases can be sequential or can be performed simultaneously but all must be addressed. It is important to note that technology, both internal and external to the clearinghouse, changes rapidly and therefore any clearinghouse must be developed to be dynamic to meet the changing nature of the technology and the changing needs of its users. Each geospatial data clearinghouse also must address the particulars of their organization such as available software, hardware, database environment, technical capabilities of staff, and the requirements of their primary clients or users. In some cases, clearinghouses have undertaken an effort to develop user requirements and assess needs prior to implementation of new services or architectures. The user needs and requirements assessment addresses all phases of the clearinghouse from both internal and external perspectives and provides the framework with which to build services and organizational capability (Kelly & Stauffer, 2000). Within the requirements phase, examination of resources available to the clearinghouse must be determined and if inadequate, acquired. There is little doubt that the key to success relies heavily on the resources of the geospatial data clearinghouse and its ability to store and provide access to large datasets and thousands of data files (Kelly & Stauffer, 2000). Another equally significant component of building an SDI is identifying how the resource will support activities in the region. The clearinghouse can bring together disparate data sets, store data for those organizations that are unable to store or provide access to their own information, and can offer access to data that crosses boundaries or regions to enable efforts that are outside traditional jurisdictions of agencies or organizations (Rajabifard & Williamson, 2001).

Metadata

The key component to any geospatial data clearinghouse is geospatial metadata. The metadata forms the core of all other operations and should be addressed in the initial phase of clearinghouse development. The Federal Geographic Data Committee (FGDC) developed its initial standards for geospatial metadata in the mid 1990's. This standard, which is used as the basis for metadata in geospatial data clearinghouses today is re-

ferred to as the Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC98). The impact of CSDGM cannot be overstated. The early metadata standard has grown over the years and been adapted and accepted internationally by the International Organization for Standardization (ISO). The metadata not only serves as a mechanism to document the data but also serves as the standard basis for distributed sharing across clearinghouses through centralized gateways or national SDI (Figure 1). The standards used to query remote catalogs have their origins in the development of the ANSI Z39.50 standard (now known as the ISO 23950 Search and Retrieval Protocol) which was originally designed for libraries to search and retrieve records from remote library catalogs (Nebert, 2004). In addition to addressing the metadata standard, a secondary yet equally important issue is format. Initially metadata was either HTML or text format and then parsed using a metadata parser into the standard fields. One of the first clearinghouses to implement XML (Extensible Markup Language) based metadata was Pennsylvania Spatial Data Access (PASDA). PASDA, which was first developed in 1996, utilized XML metadata in an effort to better manage data and to make updates and alterations to metadata more efficient (Kelly & Stauffer, 2000).

Data Acquisition

One of the most significant challenges for any geospatial data clearinghouse is the acquisition of data. Each country, state, and region face their own internal issues related to data sharing such as legal liability questions and right to know laws, therefore the data and services for each clearinghouse differ. The ideal concept for sharing would be from local government, since it can be the most detailed and up-to-date, to state and federal government clearinghouses—the hierarchy of data sharing (Figure 2). However, it can be difficult to acquire local government data unless partnerships are developed to encourage and enable sharing of local data (McDougall et al., 2005).

There are some success stories that do demonstrate the benefits of partnerships and some even with major metropolitan areas. The City of Philadelphia, which maintains one of the most advanced geospatial enter-

Figure 1. User access to remote data stores via SDI gateway

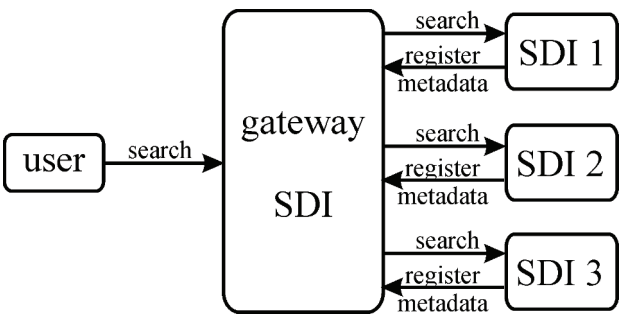


Figure 2. Traditional data sharing process from local governments to state or regional clearinghouses to national SDI gateway

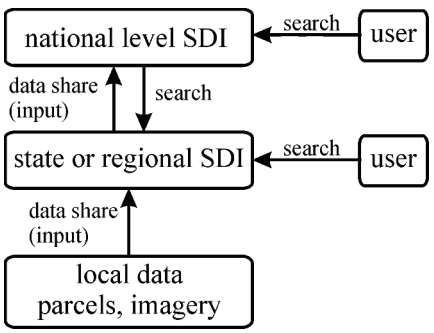


Figure 3. Internet Map Service approach to sharing data

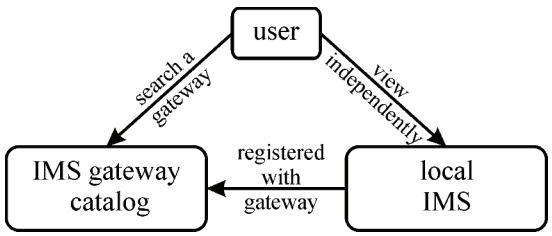
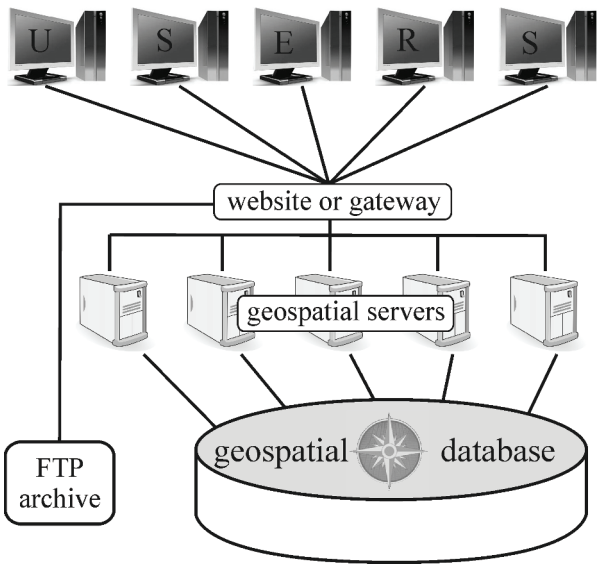


Figure 4. Common architecture of a current geospatial data clearinghouse



prise architectures in the US, shares its data through the state geospatial data clearinghouse PASDA.

However, in general the partnership approach has lead to spotty participation by local and regional organizations therefore many geospatial clearinghouses base their initial efforts on acquiring data which is freely available and which has no distribution restrictions. In the United States, this data tends to be from the Federal government. The initial data sets are comprised of elevation, scanned geo-referenced topographic maps, and aerial photographs. If the geospatial data clearinghouse is state-related, additional more detailed information such as roads, streams, boundary files, parks and recreational data are added to the data collection.

However, changes in technology have altered the concept of how data is shared. The increase in web-based GIS applications or Internet Map Services (IMS) has allowed local governments, as well as others, to share their data via IMS catalogs by registering their "service" versus sharing their data files. In this approach, access is still provided to the data, although with limited download capabilities, and local control and hosting is maintained (Figure 3).

Architecture

Geospatial data have traditionally been accessed via the Internet through file transfer protocol (FTP) or more recently viewed through web based interfaces or IMS. The earliest geospatial data clearinghouses required an FTP server, metadata that could be in HTML (Hypertext Markup Language), later XML, and simple search mechanisms driven by keywords or descriptors in the metadata. In cases such as these the onus was on the user or client to access and manipulate the data into their own environment. However, recent developments in database management and software, and advances in the overall technologies related to the Internet are

driving the rapid changes in geospatial clearinghouses. Users who are requiring more web-based services, customizing capabilities, and faster performance times combined with the exponential growth in available data has increased the requirements for clearinghouses (Kelly & Stauffer, 2000). Within the past six years, user needs have moved clearinghouse operations from simple FTP sites to complex organizations supporting large relational database management systems (RDBMS) composed of vast vector and raster data stores, specialized customization engines—reprojectors and data clipping utilities, access to temporal data, analysis tools and multiple IMS applications while still maintaining archival copies and direct FTP download capabilities (Figure 4).

EMERGING TRENDS IN GEOSPATIAL DATA CLEARINGHOSUE DEVELOPMENT

The dramatic shift from serving data and metadata through a simple web based interface to storing and providing access to thousands of unique data sets through multiple applications has placed the clearinghouse movement on the cutting edge of geospatial technologies. These cutting technologies stem from advancements in managing data in a Relational Database Management system (RDBMS) with the accompanying increase in performance of IMS and temporal data integration techniques.

RDBMS and Its Impact on Data Management

The current movement in geospatial data clearinghouses is to support and manage data within an RDBMS. In essence this approach has altered not only how the data is managed but has also improved the performance and utility of the data by enabling application developers to create IMS and other services that mirror desktop capabilities. Advancements in database technology and software such as ESRI Spatial Database Engine (SDE), in combination with databases such as Oracle or DB2, can support the management of large quantities of vector and raster data. The data is stored within the database in table spaces that are referenced, within the DB2 architecture, by configuration keywords that point to attribute table spaces and coordinate table spaces. For vector data, a spatial index grid is created which allows features to be indexed by one or more grids and stored in the database based on spatial index keys that correspond to the index grid. In general, raster data, which are in essence multidimensional grids can be stored as binary large objects or BLOBs within the table spaces. Interfacing with the database is a broker such as SDE which enables both input and retrieval of the raster BLOBs. In addition, techniques, such as the use of pyramids that render the image at a reduced resolution, increase the performance of the database for the end user. In addition, a clearinghouse that uses a RDBMS architecture is more readily able to manage temporal or dynamic data and automate processes for updating applications and services (Kelly et al, 2007). These advances in managing and retrieving data within the RDBMS/SDE environment have substantially increased performance and speed even more so in an IMS environment (Chaowei, et al, 2005). Another impact of this trend is that as relational database management become mainstreamed with the more user friendly and affordable databases such as MySQL, an open source RDBMS that uses Structured Query Language (SQL), the ability for smaller organizations to develop higher functioning IMS is within the realm of possibility.

Internet Map Services

In the past few years, IMS, which encompass everything from stand alone applications dedicated to a specific theme or dataset (i.e., "My Watershed Mapper"), have grown to represent the single most important trend in the geospatial data clearinghouse movement. The early period of IMS development included applications comprised of a set of data, predefined by the developer, which users could view via an interactive map interface within an HTML page. However, there were few if any customization capabilities and limited download capabilities. The user was simply viewing the data and turning data layers on and off. Another component of early IMS development was the use of a Web GIS interface or map as part of a search engine to access data and metadata from data clearinghouses or spatial data distribution websites (Kraak, 2004). Building on the advances of RDBMS and software, IMS developers have been able to set aside the historical constraints that hindered development in the 1990s. Initially, data used for IMS were stored in a flat

file structure, such as ESRI shapefiles with which the Internet mapping server interacted directly. While this was somewhat effective for small files in the kilobyte or single megabyte range, it became more cumbersome and inefficient for the increasing amount and size of detailed vector and raster data, such as high-resolution aerial photography. But as the use of RDBMS became the backbone of many geospatial data clearinghouse architectures, IMS developers began to take advantage of the data retrieval performance and efficiency of relational databases.

Changes in IMS have emerged over the past few years that herald a change in how users interact with geospatial data clearinghouses. The development of Web Feature Services (WFS) and Web Map Services (WMS) are pushing the envelope of an open GIS environment and enabling interoperability across platforms. There are several types of emerging IMS. These include the feature service and image service. The image service allows the user to view a snapshot of the data. This type of service is particularly meaningful for those utilizing raster data, aerial photography, or other static data sets. The feature service is the more intelligent of the two as it brings with it the spatial features and geometry of the data and allows the user to determine the functionality and components such as the symbology of the data. WFS is particularly applicable for use with real-time data streaming (Zhang & Li, 2005).

Changes in technology have allowed geospatial data clearinghouses to deploy applications and services containing terabytes of spatial data within acceptable time frames and with improved performance. IMS has moved the geospatial data clearinghouse from a provider of data to download to interactive, user centered resources that allow users to virtually bring terabytes of data to their desktop without downloading a single file.

Temporal Data

As geospatial data clearinghouses have crossed over into the IMS environment, the issue of integrating temporal data has become an emerging challenge (Kelly et al, 2007). The expectations for clearinghouses are moving toward not only providing terabytes of data but also toward providing data that is dynamic. Temporal data, unlike traditional spatial data in which the attributes remain constant, has constantly changing attributes and numerous data formats and types with which to contend (Van der Wel, et al., 2004). Temporal data by nature presents numerous challenges to the geospatial data clearinghouse because it carries with it the added dimension of time. An prime example of the complexity of temporal data integration is weather data. The development of services based on temporal information such as weather data must be undertaken with the understanding that this data is unique in format and that the frequency of updates require that any service be refreshed "on the fly" so the information will always be up to date. The number of surface points and types of data such as data taken from satellite or radar can overwhelm even a sophisticated server architecture (Liknes, 2000). However, the significance of these data to users from emergency management communities to health and welfare agencies cannot be underestimated. Therefore, it is imperative that geospatial data clearinghouses play a role in providing access to this vital data.

CONCLUSION

The changes in geospatial data clearinghouse structure and services in less than a decade have been dramatic and have had a significant impact on the financial and technical requirements for supporting an SDI geospatial data clearinghouse. As stated earlier in this section, the average cost of maintaining and expanding a clearinghouse is clear since most of the costs are assessed on personnel, hardware, software, and other apparent expenses; it is more difficult to assess the financial benefit (Gillespie, 2000). In addition, despite many changes in technology and the increasing amount of accessible data, some challenges remain. Local data is still underrepresented in the clearinghouse environment and the ever-changing technology landscape requires that geospatial data clearinghouse operations be dynamic and flexible. As trends such as IMS, the use of temporal data, and the enhancement of relational database architectures continue, geospatial data clearinghouses will be faced with providing growing amounts of data to ever more savvy and knowledgeable users.

REFERENCES

Bernard, L. (2002, April). Experiences from an implementation Testbed to set up a national SDI. Paper

presented at the *5th AGILE Conference on Geographic Information Science, Palma, Spain.*

Chaowei, P.Y., Wong, D., Ruixin, Y., Menas, K., & Qi, L.(2005). Performance-improving techniques in web-based GIS. *International Journal of Geographical Information Science*, 19 (3), 319-342.

Clinton, W.J. (1994). Coordinating geographic data acquisition and access to the National Geospatial Data Infrastructure. Executive Order 12096, *Federal Register*, 17671-4. Washington: D.C.

Craglia, M., Annoni, A., Klopfer, M., Corbin, C., Hecht, L., Pichler, G., & Smits, P. (Eds.) (2003). Geographic information in wider Europe, geographic information network in Europe (GINIE), http://www.gis.org/ginie/doc/GINIE_finalreport.pdf.

Crompvoets, J., Bregt, A., Rajabifard, A., Williamson, I. (2004). Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographical Information Science*, 18(7), 655-689.

FGDC95: Federal Geographic Data Committee, (1995). *Development of a national digital geospatial data framework*. Washington, D.C.

FGDC98: Federal Geographic Data Committee (1998). *Content standard for digital geospatial metadata.* Washington, D.C.

Gillespie, S. (2000). An empirical approach to estimating GIS benefits. *URISA Journal*, 12 (1), 7-14.

Kelly, M.C. & Stauffer, B.E. (2000). *User needs and operations requirements for the Pennsylvania Geospatial Data Clearinghouse.* University Park, Pennsylvania: The Pennsylvania State University, Environmental Resources Research Institute.

Kelly, M. C., Haupt, B.J., & Baxter, R. E. (2007). The evolution of spatial data infrastructure (SDI) geospatial data clearinghouses: Architecture and services. In *Encyclopedia of database technologies and applications* (invited, accepted). Hershey, Pennsylvania, USA: Idea Group, Inc.

Kraak, M.J. (2004). The role of the map in a Web-GIS environment. *Journal of Geographical Systems*, 6(2), 83-93.

McDougall, K., Rajabifard, A., & Williamson, I. (2005, September). What will motivate local governments to share spatial information? Paper presented at the *National Biennial Conference of the Spatial Sciences Institute,* Melbourne, AU.

National Research Council, Mapping Science Committee. (1993). *Toward a coordinated spatial data infrastructure for the nation*. Washington, D.C.: National Academy Press.

Nebert, D. (Ed.) (2004). *Developing spatial data infrastructures: The SDI cookbook* (v.2). Global Spatial Data Infrastructure Association. http://www.gsdi.org.

Rajabifard, A., & Williamson, I. (2001). Spatial data infrastructures: concept, SDI hierarchy, and future directions. Paper presented at the *Geomatics 80 Conference*, Tehran, Iran.

Stevens, A.R., Thackrey, K., Lance, K. (2004, February). Global spatial data infrastructure (GSDI): Finding and providing tools to facilitate capacity building. Paper presented at the *7th Global Spatial Data Infrastructure conference,* Bangalore, India.

Van der Wel, F., Peridigao, A., Pawel, M. , Barszczynska, M. , & Kubacka, D. (2004): COST 719: Interoperability and integration issues of GIS data in climatology and meteorology. *Proceedings of the 10th EC GI & GIS Workshop*, ESDI State of the Art, June 23-25 2004, Warsaw, Poland.

KEY TERMS

BLOB: Binary Large Object is a collection of binary data stored as a single entity in a database management system

Feature Service: The OpenGIS Web Feature Service Interface Standard (WFS) is an interface allowing requests for geographical features across the web using platform-independent calls. The XML-based GML is the default payload encoding for transporting the geographic features.

ISO 23950 and Z39.50: International standard specifying a client/server based protocol for information retrieval from remote networks or databases.

Metadata: Metadata (Greek meta "after" and Latin data "information") are data that describe other data. Generally, a set of metadata describes a single set of data, called a resource. Metadata is machine understandable information for the web.

Open GIS: Open GIS is the full integration of geospatial data into mainstream information technology. What this means is that GIS users would be able to freely exchange data over a range of GIS software systems and networks without having to worry about format conversion or proprietary data types.

SDE: SDE (Spatial Data Engine) is a software product from Environmental Systems Research Institute (ESRI) that stores GIS data in a relational database, such as Oracle or Informix. SDE manages the data inside of tables in the database, and handles data input and retrieval.

Web Mapping Services (WMS): A Web Map Service (WMS) produces maps of geo-referenced data and images (e.g. GIF, JPG) of geospatial data.

XML : Extensible Markup Language (XML) was created by W3C to describe data, specify document structure (similar to HTML), and to assist in transfer of data.