To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

Chapter 10

Interoperability and Data Integration in the Geosciences

Michael Gertz,[1] Carlos Rueda,[2] and Jianting Zhang[3]

[1] *Institute of Computer Science, University of Heidelberg, Heidelberg, Germany*
[2] *Monterey Bay Aquarium Research Institute, Moss Landing, California*
[3] *Department of Computer Science, The City College of New York, New York, New York*

Contents

10.1 Introduction

The past decade has witnessed a dramatic increase in scientific data being generated in the physical, earth, and life sciences. This development is primarily a result of major advancements in sensor technology, surveying techniques, computer-based simulations, and instrumentation of experiments. As stated by Szalay and Gray in [76], it is estimated that the amount of scientific data generated in these disciplines is now doubling every year. Organizations in government, industry as well as academic and private sectors, have made significant investments in infrastructures to collect and maintain scientific data and make them accessible to the public. Good examples of such efforts are the Sloan Digital Sky Survey in astronomy [67], the GDB Human Genome Database and Entrez Genome Database in genomics [13, 26], and the Global Biodiversity Information Facility in ecology [24], to name only a few.

More and more such domain-specific data management infrastructures are built to allow users easy access to scientific data, often in a Web-based fashion through comprehensive Web portals. However, a key challenge is to provide users with effective means to *integrate* data from diverse sources to facilitate data exploration and analysis tasks. Data integration is one of the more traditional yet still very active fields in the area of databases and data management. It is concerned with models, techniques, and architectures that provide users with a uniform logical view of and transparent access to physically distributed and often heterogeneous data sources (see, e.g., [5, 11, 72, 87]). Data integration is a key theme in many e-commerce and e-business IT infrastructures, often called enterprise information integration [36]. In these application domains, the objective is to integrate business and consumer data from different transactional databases in order to obtain new information that drive business activities and decisions. Nowadays, several commercial and open-source data integration platforms exist that help businesses to integrate (typically relational) data from transactional databases, leading to data warehouse and federated database architectures.

It seems natural to apply similar techniques realized in those business-oriented data integration platforms to scientific data collections as well. However, because of the complexity, unprecedented quantities, and diversity of scientific data, traditional schema-based approaches to data integration are in general not applicable. In many scientific application domains, there often is no single conceptual schema that can be developed from the data and schemas associated with the individual data sources to be integrated. Furthermore, scientific data integration often occurs in an ad hoc fashion. For example, data relevant to evaluate a scientific hypothesis needs to be discovered and dynamically integrated into often complex data analysis and exploration tasks without requiring to persistently store the data used in these tasks. The problem many scientists are facing nowadays is how to easily make use of the ever-increasing number of data repositories in an effective way.

A prominent domain where these problems become more and more apparent and pressing is in the geosciences. Geospatial data, that is, data that is spatially referenced to Earth, have become ubiquitous. This is primarily due to major advancements in remote-sensing technology, surveying techniques, and computer-based simulations. As an example, the satellites operated by NASA and NOAA generate dozens of terabytes of imagery and derived data products per day, leading to one of the fastest growing repositories with petabytes of science data. In the year 2003, NOAA already maintained about 1,300 databases containing more than 2,500 environmental variables [1]. The diverse types of geospatial data collected by federal and local governments as well as organizations in industry and academia play a significant role in developing mission-critical spatial data infrastructures [32, 52].

The use of geospatial data obtained through observations and simulations and their management in spatial data infrastructures have become essential in many application domains. These include environmental monitoring, climate research, disaster prevention, natural resource management, transportation, and decision support at various levels of local and state governments. The types of geospatial data considered in these domains come in a variety of types. Common types include maps and imagery from air and space-borne instruments, vector data describing geographic objects and features, outputs from simulations, and numerous types of real-time sensor data. In particular, the latter are an emerging data source, driven by large-scale environmental observation networks such as envisioned by NEON [48].

With such a proliferation of a wide range of geospatial data repositories, many of which are readily accessible through the Web, it is imperative to achieve a high degree of interoperability among these systems as a prerequisite to facilitate data-integration tasks. By realizing this objective, geospatial data that is managed in specialized repositories in support of specific domains and tasks can serve whole communities and scientists in different disciplines.

In this chapter, we present the current trends and technologies in support of developing interoperable geospatial data sources and management architectures that *enable the efficient sharing, use, and integration of physically distributed and heterogeneous geospatial data collections*. Our primary focus is on emerging technologies that facilitate true interoperability among geospatial data repositories, such as the development and implementation of standards for geospatial content and services promoted by the Open Geospatial Consortium (OGC) [78]. A key concept underlying this approach is (geospatial) Web services, which realize a standard way to interoperate with diverse geospatial data management infrastructures and to access heterogeneous forms of geospatial data in a uniform and transparent fashion.

Such type of interoperability, of course, is only one ingredient to effective data-integration approaches. Compared with data-integration techniques for traditional relational databases, there are several special properties pertinent to geospatial data. For example, a complicating factor in integrating geospatial data is the variety of formats in which the data is managed, ranging from flat files to specialized geographic information systems (GIS). As we will illustrate in the following sections, geospatial Web services provide an effective means to request geospatial data from heterogeneous repositories in a format suitable for data integration tasks. Such services help greatly in dealing with data heterogeneity and conflict resolution aspects in data integration. Further data integration challenges, such as heterogeneity of the data in terms of structure and semantics are often dealt with by employing standard representation formats and taxonomies, respectively. In particular, for these two aspects, we show that there have been significant achievements in various subdisciplines of the geosciences, especially in the development of schema frameworks for describing geospatial data and metadata/taxonomy frameworks that focus on the semantics of geospatial data components.

A novel aspect we focus on in this chapter is the integration of streaming real-time data, which is becoming a predominant source of geospatial data, for example, in remote-sensing and sensor observation networks. We describe recent technologies that have been developed for the service-based management and consumption of streaming geospatial data and show how computing infrastructures can be built that effectively consume diverse static as well as dynamic (streaming) geospatial data from heterogeneous and distributed data sources. General techniques for managing streaming data are discussed in Chapter 11.

The remainder of this chapter is organized as follows. In Section 10.2, we review basic geospatial data management and integration concepts, including data formats, metadata standards, and existing approaches and techniques to geospatial data integration. In Section 10.3, we discuss in detail the technologies surrounding geospatial Web services. Furthermore, we outline emerging technologies in the context of sensor Web enablement architectures. In Section 10.4, we use a relevant practical scenario from the environmental sciences to demonstrate how the different techniques presented in this chapter can effectively be deployed to perform geospatial data integration tasks, including the integration of real-time sensor data. We conclude the chapter in Section 10.5 with a summary.

10.2 Geospatial Data Management and Integration

In this section, we review some fundamental concepts and techniques for the management of geospatial data. In Section 10.2.1, we give a brief overview of geospatial data models and representation formats. Section 10.2.2 outlines some standard approaches to managing geospatial data. Our particular focus in these two sections is on issues relevant to interoperability and integration aspects. After a discussion of schema and metadata concepts in Section 10.2.3, we discuss in Section 10.2.4 some existing approaches for integrating geospatial data.

10.2.1 Geospatial Data Models and Representations

Depending on the application domain and collected geospatial information, geospatial data can be modeled and represented in different ways. The two most common approaches to model geographic information are using either an *object-based model* or a *field-based model* (see, e.g., [64, 70]). In an object-based model, geographic objects correspond to real-world entities (also called *features*) about which information

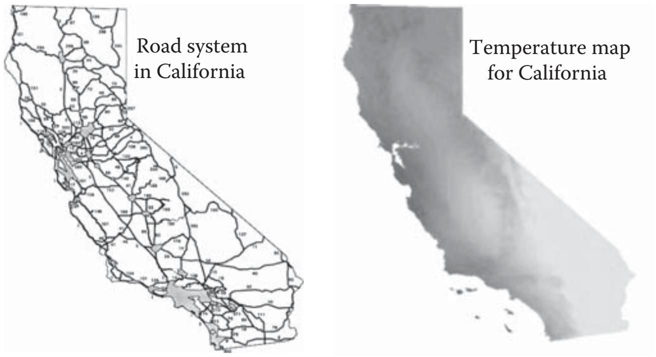

Figure 10.1 Examples of object-based (left) and field-based (right) geospatial data representation.

needs to be managed. A feature typically has two parts: (a) a spatial component (or *spatial extent*), which specifies the shape and location of the object in the embedding space; and (b) a descriptive component that describes the *nonspatial properties* of the feature in the form of attributes. The spatial extent of an object is typically modeled as a point, polyline, or polygon, depending on the required spatial granularity and scale of the data to be managed. For the representation of a collection of features, different approaches exist, such as the network model, spaghetti model, or topological model [64]. The left part of Figure 10.1 shows an example of an object-based presentation of geographic information (a road network).

In field-based approaches, the space to be modeled is partitioned (tessellated) into two- or multidimensional cells, a cell having a spatial extent. With each cell one or more attribute values are associated, each attribute describing a continuous function in space. A typical example of field-based data are multispectral or hyperspectral raster imagery obtained from remote-sensing instruments. Field-based data are also common as outputs of simulations where with each point in space a set of attribute values (measurements) is associated. Note that in a field-based model, there is no notion of objects but observations of phenomena in space, which are described by attribute values (measurements) that vary with the location in space. The right part of Figure 10.1 shows an example of a field-based representation (estimated temperature over an area).

In order to precisely describe the spatial extent of geographic objects or cells in a raster image, it is important to have a *spatial reference system (SRS)* (or *coordinate reference system (CRS)*) underlying the space in which features and phenomena are modeled. A reference system is a particular *map projection* that represents the two-dimensional curved surface of the Earth. There are numerous such map projections used in practice, ranging from global projections such as latitude/longitude or Universal Transverse Mercator (UTM) to parameterized local ones tailored to specific regions on the Earth's surface, such as the State Plane Coordinate System [74] used in the United State From a spatial data management point of view and in particular for the integration of diverse datasets, an important aspect is to be able to re-project geospatial data from one reference system to another one [39, 86].

10.2.2 Geospatial Data Management Systems and Formats

Compared with data management systems for relational data, which are all based on the same model (the relational model) and make use of the same language (SQL), there is a plethora of commercial and open-source systems for managing geospatial data. In the following, we give a brief overview of the different types of systems and focus on aspects that are relevant to data integration approaches.

GISs are the predominant type of systems to manage, store, analyze, and display geographic data and associated attributes that are spatially referenced to the Earth [45]. A widely used type of GIS is ESRI's (Environmental Systems Research Institute's) ArcGIS products, such as ArcView to view spatial data, create maps, and perform basic spatial analysis operations. ArcInfo is an advanced version of the ArcGIS product line that also includes functions for manipulating, editing, and analyzing geospatial data [17, 60] and services for geoprocessing and geocoding [16]. ArcGIS also provides different types of Web services to access geospatial data.

There are also traditional relational database management system vendors that offer spatial extensions to their relational engines. For example, Oracle Spatial provides several functions for storing, querying, and indexing spatial data, including raster and gridded data [46]. The spatial extension models a majority of the spatial types and operations described in the SQL/MM spatial standard [75]. IBM's DB2 product line also offers spatial extensions to their relational DB2 core system such as the DB2 Spatial Extender and the DB2 Geodetic Extender [38]. Also here, the spatial extensions implement types and functions specified in the SQL/MM standard.

Please provide full term.

Prominent open-source GIS type systems are PostGIS [62], the spatial extension of the object-relational database management system PostgreSQL, and the Geographic Resources Analysis Support System (GRASS) [33, 51]. Like the spatial extensions for Oracle and DB2, PostGIS follows the *Simple Features for SQL specification* developed as an implementation specification by the OGC [55]. This standard specifies the storage of different types of geographic objects (points, lines, polygons, etc.) and includes specifications for various spatial operators to derive new objects from existing ones. PostGIS makes use of the proj.4 library [63] for converting geographic data between different map projections, an important functionality to integrate geospatial objects that are based on different reference systems.

GRASS provides a variety of functions to manage raster data and topological vector data. It natively uses and supports a number of vector and raster formats, which are expanded with several other formats using the Geospatial Data Abstraction Library (GDAL) [25]. GRASS offers the option to manage nonspatial attributes associated with geographic objects and raster images in either files or an SQL-based database management system.

Besides the above GIS type of data management infrastructures, geospatial data are also often managed just at the file level. That is, applications generate geospatial data and simply record them in standard file formats for consumption by and exchange

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

with other programs. One can basically distinguish between file formats for vector data (object-based data) and file formats for raster or gridded data (field-based data). One of the most common formats for vector data are *shapefiles*, which have been developed by ESRI and are used to exchange data among ESRI products and other software [18]. Another important, although less widely used, format for vector data is the Topologically Integrated Geographic Encoding and Referencing (TIGER) format used by the U.S. Census Bureau. It is employed for modeling geographic information such as roads, rivers, lakes and census tracts [82].

For raster and gridded data, widely used file formats are the Network Common Data Form (NetCDF) [49], the Hierarchical Data Format (HDF5) [37], and GeoTIFF [50]. These file formats only represent a small but important portion of a large collection of scientific data formats (many of which also come in an XML framework) that have been developed over the past decades in different disciplines.

The above discussions about the variety of commercial and open-source geospatial data management software as well as file formats for the exchange of complex (geo)spatial data clearly illustrate that achieving interoperability among heterogeneous geospatial data sources is a great challenge.

10.2.3 Schemas and Metadata

An essential ingredient to any data integration approach is to have information about the schemas as well as metadata for schema components and the data managed in heterogeneous scientific data repositories. In the following, we first discuss an emerging standard for geospatial data to represent both schema information and data and then detail some prominent metadata frameworks used in the context of geospatial data.

10.2.3.1 GML Application Schemas

The Geography Markup Language (GML) is an XML-based specification developed by the OGC for representing geographic features [42, 56]. GML serves as an open interchange format for geospatial data as well as a modeling language for geographic information. In GML, real-world objects are called *features* and have a spatial component (geometry) and nonspatial properties. The most recent GML version, 3.1, is being standardized as ISO 19136. While earlier GML versions used Document Type Definitions (DTDs), the later versions are based on XML-Schema. GML version 3.x also includes support for two-dimensional complex geometries and topology, three-dimensional geometries, spatial and temporal reference systems, and visualization.

Should this say version 3.1?

Because GML is based on XML-Schema, it allows users to create their own application schemas by making use of GML (core) schema components such as geometry, topology, and time, and follow the simple, structured rules of the GML encoding specification. GML application schemas are very flexible in that they allow users to tailor and extend predefined GML data types (mostly geometrical and topological) to specific needs in an application domain. GML also serves as data exchange format for geospatial data, an aspect that is particularly important to achieve a high degree of interoperability among geospatial data repositories through geospatial Web services.

10.2.3.2 Metadata Standards

There are many metadata frameworks for spatial data and applications that have geospatial components. Most of these frameworks and initiatives are driven by individual science communities. Metadata frameworks can be found at all data management levels, ranging from metadata associated with traditional database and GIS schemas to approaches where metadata is simply encoded as part of a file format containing the (geo)spatial data.

The most widely used geospatial metadata standard in GIS products is the standard developed and maintained by the Federal Geographic Data Committee (FGDC) [21]. The FGDC developed the Content Standard for Digital Geospatial Metadata (CSDGM) in 1994, which is often simply referred to as the FGDC metadata standard [20]. This standard has components to describe the availability, fitness for use, and access and transfer information of geospatial datasets. According to the CSDGM version 2 published in 1998, Section 1 has entries to describe the geographical area a geospatial dataset covers; Section 3 describes the spatial data model that is used to encode the spatial data (vector/raster) or other possible methods for indirect georeferencing; and Section 4 describes the information about the spatial reference system.

In addition to the CSDGM, several other metadata standards have been developed over the past few years for different application domains in the geosciences and environmental sciences. For example, the Ecological Metadata Language (EML) developed by the National Center for Ecological Analysis and Synthesis (NCEAS) has been widely adopted in the ecological data management community [10, 12, 40]. EML has been designed as a collection of modules and has an extensible architecture. For the data module, EML has detailed structures to describe tabular, raster, and vector data. In EML, the metadata is much more tightly coupled with data, compared with that of the FGDC metadata standard. Such a coupling is an important aspect in metadata-driven data integration [40]. Another extensive data description framework for Earth science initiatives is the Semantic Web for Earth and Environmental Terminology (SWEET) developed by NASA [47]. SWEET is a standard vocabulary rather than a full-fledged metadata framework, and it includes a variety of data description components in the context of physical phenomena; processes; and properties, sensors, space, time, and units.

10.2.4 Approaches to Integrating Geospatial Data

Traditional data integration basically follows a schema-matching approach in which related schema components (relations and attributes) from the different sources are identified, homogenized, and suitably integrated to provide the user with a single conceptual view over the data managed at the sources (see, e.g., [5, 11, 72, 87]). Using such a view, the distributed data then can either be physically integrated at a single site or queried in a uniform and transparent fashion. The former approach then leads to some kind of data warehouse that physically stores the integrated data, but now in a homogeneous representation and format, leading to the *physical integration* of data. The latter approach, on the other hand, results in a federated or multidatabase

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

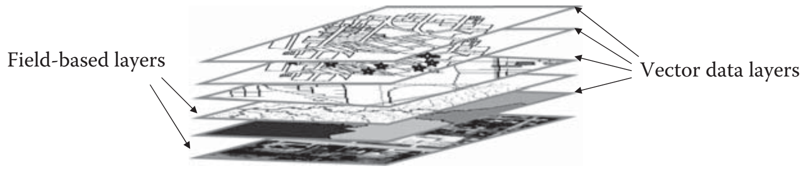

Figure 10.2 Illustration of an overlay of themes in a GIS. Geo-referenced and aligned layers include both vector data and field-based data.

system, realizing a so-called *logical integration*. Key to the integration is resolving the various types of structural and semantic heterogeneities that occur due to differences in data representation and meaning [71].

For integrating geospatial data sources, the approach can be significantly more complicated, especially because there is a wider variety of geospatial data types (compared to just relational data), including various vector data representations and formats for field-based data, as discussed above. But, what is actually meant by *integrating geospatial data*? In practice, the most common view of this is to have a GIS that allows users to overlay different *themes* (or *layers*). That is, for a given geographic area, there are several georeferenced themes that represent different characteristics of that area. Figure 10.2 illustrates an example in which several themes based on vector and field-based data are overlayed. Theme overlays allow to view and explore geographic data in different contexts. Being able to visualize data in context is an important functionality in integrating diverse types of geospatial data.

A theme can be represented by either vector data or field-based data. A road network with roads being individual features, for example, would be represented as vector data, whereas a vegetation index would be represented as field-based data (more specifically a raster image). If the data for the layers come from different sources, two problems can occur. First, the system used to integrate the data has to be *interoperable* with the other systems the geospatial data are retrieved from. Here, *interoperability* means that systems can exchange information and data using standard protocols and formats. As we will discuss below, a high degree of interoperability can be achieved when distributed and heterogeneous geospatial data sources can be uniformly accessed using geospatial Web services.

Second, the geospatial data may come in different formats with conflicting structures and semantics. For example, if two sources provide vector data for the same theme and region, the data might conflict in terms of their spatial components as well as their descriptive components (see Section 10.2.1). Such a situation can even occur if both datasets are based on the same projection (spatial reference system), have been georeferenced/aligned and have the same scale (spatial resolution). Re-projection, georeferencing, and scaling are tasks that are frequently used in the context of remotely sensed imagery and are typically performed on the datasets prior to their overlay or integration. Another typical example often occurring in practice is when some raster imagery is overlayed with vector data. Phenomena in the image might not align or match up with the features modeled by the vector data. Approaches to resolving these types of conflicts are known as *conflation*, meaning to "replace two or more versions of the same information with a single version that reflects the pooling, weighted averaging, of the sources" [45].

The key in dealing with conflicting spatial components of two or more datasets to be integrated is to make use of the location information associated with geospatial objects (and cells/pixels in a raster image), something unique to spatial datasets. Several approaches have been proposed that deal with the integration of vector data and road maps in particular and the combination of imagery with vector data [6, 66, 83]. More fine-grained approaches have been developed for finding corresponding objects in datasets to be integrated. Corresponding objects (features) represent the same real-world entity but are possibly misaligned across different data sources. Approaches for point-based data have been presented by Beeri et al. in [2, 3], referred to as *location-based join*. Related approaches are so-called entity resolution techniques, which try to determine the true location of a real-world entity in case geospatial data about the entity comes from a collection of data sources [69].

For resolving spatial conflicts, that is, if the same real-world entity has conflicting feature location information in the different sources, nonspatial attributes associated with the features can help in resolving such conflicts. For example, if it is known that a feature has been updated recently at a source, the feature at this source might be more likely to represent the correct location information about the real-world entity. As with any other data integration approach, the quality of the data plays a crucial role in resolving individual spatial and nonspatial data conflicts [73].

Once different features have been matched to the same real-world entity, the next step is to resolve conflicts that might exist among the descriptive attributes. As these are ordinary attributes such as in relational databases, respective approaches can be used. In the context of geospatial data such attributes are typically based on metadata standards and application schemas described in Section 10.2.3, which are likely to produce a more coherent data description in terms of semantics.

10.3 Service-Based Data and Application Integration

In the following, we present emerging standards, techniques, and architectures that enable interoperability among distributed and heterogeneous geospatial data sources. In Section 10.3.1, we outline the relationships between interoperability and data integration aspects. An overall framework for data integration employed by the techniques presented in this chapter is the service-oriented architecture, which is described in Section 10.3.2. In Sections 10.3.3 and 10.3.4, we give an overview of service registry and geospatial Web services, respectively. We place a particular focus on services that deal with real-time sensor data, described in Section 10.3.5. We conclude the section with a brief overview of a practically relevant alternative to geospatial Web services.

10.3.1 Approaching Integration through Interoperability

Interoperability among heterogeneous and distributed data sources is a fundamental requirement not only in the context of scientific data management, but in any type of distributed computing infrastructure. *Interoperability* is generally defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [53]. Interoperability can be achieved at different levels of network protocols and data exchange formats. In several scientific application domains, interoperability among data repositories and applications has become a main driver to facilitate scientific data management and exploration on a large scale. Grid computing infrastructures have significantly contributed to this development [22] and are widely employed in science domains, such as in Earth observation [27], climate modeling [15], and physics [35], to name only a few. A more recent trend in these science initiatives is to increase interoperability aspects through *service-oriented science* [23]. A well-known early example that realizes such an approach is the WorldWide Telescope [34].

One major driver in the area of geoprocessing and geospatial data management technologies is the OGC Interoperability Institute [54], where the OGC is also developing and promoting diverse types of geospatial Web services. Such type of Web services play an increasing role in geospatial data integration frameworks. Services do not just provide easy access to diverse types of geospatial data using standard protocols and interfaces, but they also often offer functionality that helps in resolving data conflicts. For example, requesting data in a particular projection or at a particular scale are important data preprocessing steps that already can be accomplished by services rather than at the data integration site. In this sense, such services provide some application functionality too. There are a few geospatial Web services approaches that address both interoperability and integration aspects, for example, based on mediation [9], services [44], or a combination of service and mediation-based techniques [4, 19, 43]. In the following, we describe how integration infrastructures can be built based on such services and architectures.

10.3.2 Service-Oriented Architectures

In the past few years, there have been significant developments in terms of architectures and standards that help developers build Web-based services that allow for a uniform and transparent access to data managed at different sources. One such development is the service-oriented architecture (SOA) (see, e.g., [14]), which allows an effective cooperation among data sources and data processing components hosted at different organizational units. In particular, SOA supports reusability and interoperability of software and service components on the Web, thus increasing the efficiency of developing and composing new services. In a SOA-based system, all data and process components are modeled as Web services.

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

10.3.3 Registry and Catalog Services

Of particular interest in this chapter are catalog and registry services. As scientific data are accumulating in an ever increasing speed, it is very difficult if not impossible for users to know exactly the details of all the data that might be relevant to their project. As such, repositories that provide catalog services and allow users to interactively or programmatically search and retrieve metadata that are related to the use of the datasets are playing an inreasingly important role in scientific data management. In the context of geospatial applications, OGC's Catalog Service for the Web (CSW) Implementation Standard [58] provides this functionality in the form of several operations: the mandatory GetCapabilities operation returns metadata about the specific repository server (ServiceIdentification), the operations supported by the service including the URL(s) for operation requests (OperationMetadata), the type of resource cataloged by the repository server (Content), and the query language and its functionality supported by the repository. The GetRecords operation allows users to specify query constraints and metadata to be retrieved and returns the number of items in the result set and/or selected metadata for the result set. The DescribeRecord operation allows a client to discover elements of the information model supported by the target catalog service. The optional GetDomain operation is used to obtain runtime information about the range of values of a metadata record element. Finally, the mandatory GetRecordByID request retrieves the default representation of catalog records using their identifier. Through the GetCapabilities → GetRecords → DescribeRecord → GetDomain → GetRecordByID sequence, users are able to probe the repository server's capabilities, search the repository, negotiate the format of the metadata and finally retrieve the metadata of the dataset(s) of interest.

10.3.4 Geospatial Web-Services and Standards

The OGC was founded with the mission of advancing the "development of international standards for geospatial interoperability" [78]. The OGC currently comprises, at time of writing, over 350 companies, universities, and government agencies from around the world. In the Earth sciences in particular, the role of standard data and interface protocols is crucial in the context of climate monitoring and forecasting. The National Weather Service [41], for example, has recently started to make forecast data available to users using Web Feature Service (WFS) and Geography Markup Language (GML), two of the open standards developed by OGC.

In this Web service framework, the concepts of coverage, feature, and layer play a key role in publishing and accessing diverse types of geospatial datasets through OGC Web services. Both coverage and feature provide associations between observed or measured values with a geographical domain, such as a particular region or spatial extent (see also Section 10.2.4). A coverage can be thought of as a measurement that varies over space, while a feature is a spatial object that has associated measurements. In the context of remotely-sensed data, for example, a satellite image covering an area can be represented as a coverage. On the other hand, the observation values of a

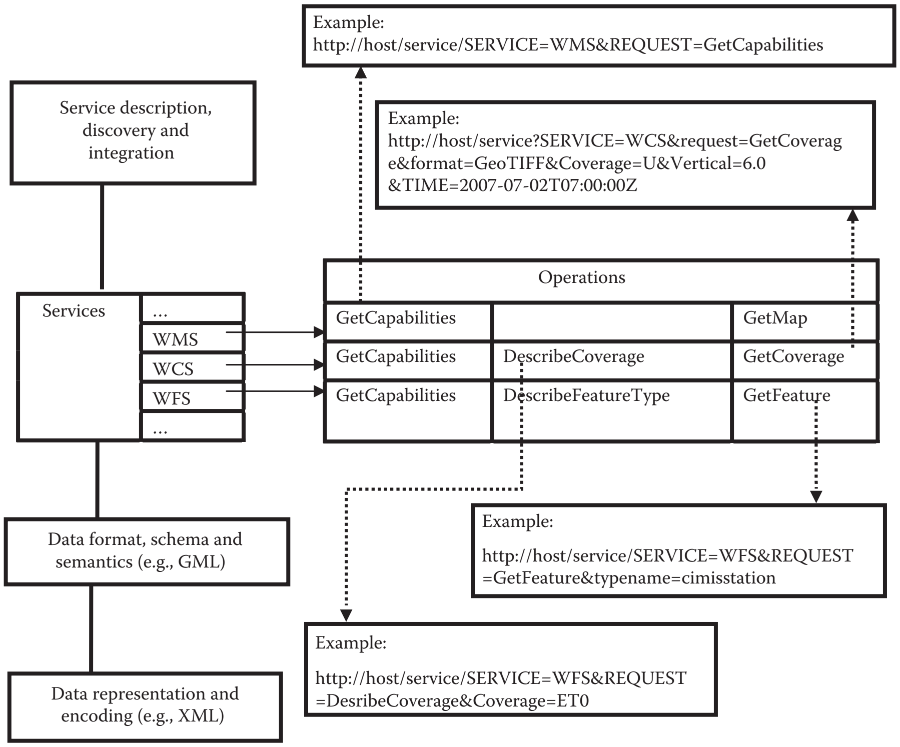
Example:
http://host/service/SERVICE=WMS&REQUEST=GetCapabilities
Service description, discovery and integration
Example:
http://host/service?SERVICE=WCS&request=GetCoverag e&format=GeoTIFF&Coverage=U&Vertical=6.0 &TIME=2007-07-02T07:00:00Z
Operations
Services
...
WMS
WCS
WFS
...
GetCapabilities
GetMap
GetCapabilities
DescribeCoverage
GetCoverage
GetCapabilities
DescribeFeatureType
GetFeature
Example:
http://host/service/SERVICE=WFS&REQUEST =GetFeature&typename=cimisstation
Data format, schema and semantics (e.g., GML)
Example:
http://host/service/SERVICE=WFS&REQUEST =DesribeCoverage&Coverage=ET0
Data representation and encoding (e.g., XML)

Figure 10.3 OGC Services, Operations, and Example Calls (indicated by dotted lines) for Web Map Service (WMS), Web Coverage Service (WCS), and Web Feature Service (WFS).

(point-based) weather station can be represented as a feature. A layer, which basically corresponds to the concept of a theme, can be either a gridded coverage or a collection of similar features.

The ability to map heterogeneous forms of geospatial datasets to a few simple types (such as features, coverages, and layers) greatly reduces the complexity of diverse data types in application and data integration scenarios in particular, and it makes it possible to standardize publishing datasets using Web services. Although it is beyond the scope of this chapter to give a detailed technical description of OGC Web standards, Figure 10.3 shows the three major OGC standard services along with example operations. These services, which cover the two different types of geospatial data (features and coverages) and their visualization, are as follows:

- *Web Feature Service (WFS)*: WFS defines interfaces for querying and retrieving features based on spatial and nonspatial properties of the features. The data is exchanged between a Web Feature Service and the client in the form of GML documents, which in the case of this service encode vector data.

- *Web Coverage Service (WCS)*: WCS defines interface to query and retrieve spatially referenced coverages, i.e., gridded or raster data.
- *Web Map Service (WMS)*: This service produces maps (in the form of digital images) of spatially referenced data (i.e., features or coverages) from a data source managing geographic information. Standard image formats that can be requested by a client include PNG, GIF, GeoTIFF, and JPEG.

In summary, WFS is used for object-based data, and WCS is used for gridded/raster data. Displaying such data and their overlays is done using WMS. To illustrate the functionality of these OGC services, consider the WCS standard as an example. Like the other two services, it defines a mandatory GetCapabilities operation, which allows clients to get WCS server metadata, including an optional list of the offered coverages with some metadata for each coverage. In addition, WCS also defines a mandatory DescribeCoverage operation that allows clients to get more metadata about identified grid coverage(s), including details about the spatial extent of the coverage. A WCS GetCoverage operation requests and returns coverages representing space-time varying phenomena. In general, through a sequence of GetCapabilities → GetMap (WMS), GetCapabilities → DescribeCoverage → GetCoverage (WCS), and GetCapabilities → DescribeFeatureType → GetFeature (WFS), client applications are able to retrieve both metadata and data subsets of interests in a standard way.

The realization of the above services typically occurs in the form of middleware layers that clients can access through the Web. Among the most prominent representatives of such middleware layers are the open source systems GeoServer [29] and MapServer [81]. Either system provides a client with transparent access (using the above OGC services) to diverse types of data stores. That is, these servers can be configured to access geographic data managed in, for example, PostGIS, Shapefiles, or Oracle Spatial, and to provide clients with access to the data through WFS, WMS, and WCS interfaces. In this sense, such a type of middleware layer already realizes an important component to data integration scenarios, namely the transparent and uniform access the diverse geospatial data sources. For example, using the services, one can request data in a particular (common) coordinate system. Thus, the services help in resolving some data heterogeneity issues.

10.3.5 Sensor Web Enablement

Of particular relevance are the activities recently taken by the OGC Sensor Web Enablement (SWE) program, one of the OGC Web Services initiatives [57, 61]. The SWE initiative seeks to provide interoperability between disparate sensors and sensor processing systems by establishing a set of standard protocols to enable a "Sensor Web," by which sensors of all types in the Web are discoverable, accessible, and taskable. The SWE standards allow the determination of the capabilities and quality of measurements from sensors, the retrieval of real-time observations in standard data formats, the specification of tasks to obtain observations of interest, and the asynchronous notification of events and alerts from remote sensors.

SWE components include models and XML Schemas (SensorML, Observations & Measurements, TransducerML) and Web service interfaces (SOS, SPS, SAS, WNS), which are briefly described as follows (see [57, 61] for more details):

- SensorML, *Sensor Model Language*: An XML Schema to describe sensors and sensor platforms. SensorML provides a functional description of detectors, actuators, filters, operators, and other sensor systems.
- O&M, *Observations & Measurements*: A specification for encoding observations and measurements from sensors.
- TransducerML, *Transducer Markup Language*: A specification that supports real-time streaming of data to and from transducers and other sensor systems. Besides being used to describe the hardware response characteristics of transducers, TransducerML provides a method for transporting sensor data.
- SOS, *Sensor Observation Service*: This Web service interface is used to request and retrieve metadata information about sensor systems as well as observation data.
- SPS, *Sensor Planning Service*: Using this Web interface, users can control taskable sensor systems and define tasks for the collection of observations and the scheduling of requests.
- SAS, *Sensor Alert Service*: Through this Web service interface, users are able to publish and subscribe to alerts from sensors.
- WNS, *Web Notification Service*. This Web service interface allows the asynchronous interchange of messages between a client and one or more services (e.g., SAS and SPS).

In accordance with the philosophy of Web services in general, and the SWE initiative in particular, data consumers should be concerned only with registries and service interfaces. For example, an SOS provider needs to be "discovered" first through a registry mechanism, which is the OGC Catalog Services (see Section 10.3.3) in the SWE context. Section 10.4 describes specific elements from SensorML, O&M, and SOS that have been included in a prototype to chain data stream processing services. The general sequence of steps to obtain sensor metadata and data is shown in Figure 10.4.

As an SOS service, the provider first provides a *capabilities* document as a response to a GetCapabilities request by a client. This document includes the identification of the provider and the description of the offered services, that is, the available streams in the system, which are organized in the form of *observation offerings*. An offering includes information about the period of time for which observations can be requested, the phenomena being sensed, and the geographic region covered by the observations. A schematic example of a capabilities document is shown in Figure 10.5. (We use a simplified structure style to illustrate XML documents in this section; we use ◇ symbols to indicate relevant XML elements, and example values are shown in cursive.)

Once a client is interested in a particular geospatial data stream, it will submit a DescribeSensor request to the provider. The response is a document describing the sensor that generates the data stream. This response takes the form of SensorML or a

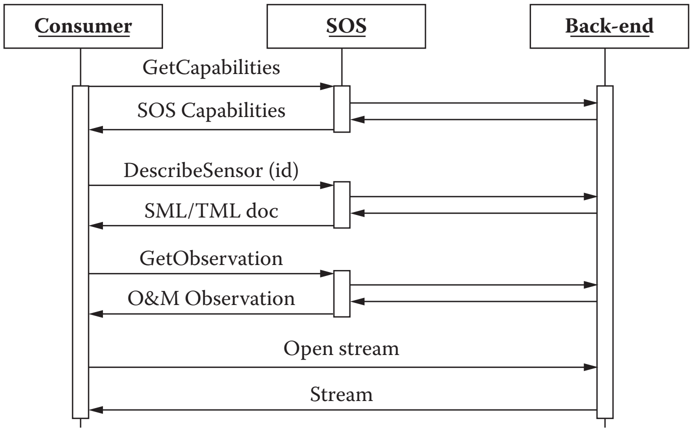

Figure 10.4 SOS sequence diagram for getting metadata and observations, including access to streaming data

TransducerML document. Next, the client will request the actual data from the sensor. This is done by submitting a GetObservation request. The corresponding response is an O&M document containing the observation, either explicitly (inlined in the document), or by providing a hyperlink to the actual data. This is illustrated in the lower part of Figure 10.4 where the client requests a connection to the data stream

- ◇ sos:Capabilities
    - ◇ ServiceIdentification
        - ◇ Title, ◇ Abstract, ◇ Keywords
    - ◇ ServiceProvider
        - ◇ Name, ◇ ContactInfo, ◇ Site
    - ◇ Operations
        - *GetCapabilities, DescribeSensor, GetObservation*
    - ◇ FilterCapabilities
        - ◇ Spatial, ◇ Temporal, ◇ Scalar
    - ◇ ObservationOfferings
        - *visible radiance, near-infrared, mid-infrared*
        - ◇ boundedBy
            - ◇ Envelope: *field of view of GOES Imager*
        - ◇ TimePosition
            - ◇ beginPosition: *2003-10-10*
            - ◇ endPosition: *indeterminatePosition*
        - ◇ Result format: *text/xml;subtype="OM"*

Figure 10.5 Schematic example of an SOS capabilities document with observation offerings exemplified with typical spectral sensors from an environmental satellite (NOAA's GOES satellite).

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

directly to the back-end system. We use this mechanism for allowing the access to real-time data in an application scenario, detailed in Section 10.4.

10.3.6 OPeNDAP

We conclude this section by giving a brief overview of a framework that also provides services, mostly in the form of protocols, to manage and access scientific data. While OGC standard components are designed to handle georeferenced data, the Open Source Project for Network Data Access Protocol (OPeNDAP) standards describe the management of multidimensional array data that are not necessarily georeferenced [59]. OPeNDAP includes specifications for encapsulating structured data, annotating the data with attributes and adding semantics that describe the data. In addition to the Distributed Oceanographic Data System (DODS) protocol that allows users to transparently access distributed data across the Internet in a way similar to the GetCoverage operation in the OGC WCS standard, OPeNDAP has protocols for exchanging metadata. More specifically, the dataset attribute structure (DAS) is used to store attributes for variables in the dataset. The dataset description structure (DDS) is a textual description of the variables and their classes that make up a scientific dataset.

Existing implementations based on OPeNDAP standards provide a convenient framework for retrieving multidimensional scientific data using simple HTTP-GET requests and are widely used by governmental organizations such as NASA and NOAA to serve satellite, weather, and other Earth science data [65, 84]. Since there are no coordinate referencing systems involved in OPeNDAP standards, they are best used for datasets with a common underlying coordinate system. However, OPeNDAP services may not be sufficient when datasets with different coordinate systems need to be integrated. In such cases, OGC-based services are more suitable. OGC standards and OPeNDAP standards are not necessarily exclusive. By enhancing multidimensional arrays data with proper coordinate systems, it is possible to construct coverages and serve the data using OGC WCS and WMS standards.

10.4 An Example of an Integration and Interoperability Scenario

The concept of service-based geospatial data integration can be demonstrated in many environmental monitoring scenarios. In this section, we consider a particular environmental scenario involving several integration aspects and describe the features an integration framework should in general provide to support such kind of scenarios.

10.4.1 Environmental Modeling Task — Evapotranspiration

We elaborate on a particular use case involving the integration of real-time evapotranspiration observations and its comparison with estimations from a weather model for accuracy assessment. Evapotranspiration (or ET) is a term used to describe the sum of evaporation and plant transpiration from the Earth's land surface to the atmosphere, which is an important part of the water cycle. ET estimations are used in irrigation scheduling, watershed management, weather forecasting, and the study of long-term effects of land use change and global climate change [31]. A standard reference evapotranspiration, denoted ETo, can be determined by using meteorological measurements, which can be obtained from multiple sources for the same region. Compared with station-observed ETo (point data), the weather model–based ETo (raster data) has a continuous spatial coverage. In general, the model output accuracy needs to be verified against the station observations. Assuming the observed and the predicted data are published as WFS and WCS services, respectively, data integration is needed to retrieve the model predicted data at the station locations and compare the values, possibly after normalization and unit conversions.

The scientific goal in this scenario is the visualization, monitoring, and validation of model-based evapotranspiration for different eco-regions and selected locations in California. This requires the overlay of ETo from the various sources on a single display for visual analysis. We show next a geospatial integration approach that allows the realization of this scenario.

10.4.2 Integration Platform

The overall conceptual architecture for geospatial data integration and interoperability, which consists of data sources, structured repositories, service middleware systems, and client applications, is depicted in Figure 10.6. Data sources include remotely sensed imagery, model outputs, GIS stores, sensor network systems, as well as other service-enabled data providers. The structured data repositories in general refer to data management systems including databases and data stream engines. Service middleware refers to service enabling infrastructures that make the data available to clients by means of Web service standards. Client applications allow users to search, query, and retrieve metadata and data by using the provided Web services (CSW, WFS, SOS, etc.), possibly in combination with more traditional access mechanisms (HTTP, FTP, etc.) with the back-end data repositories.

In general, a service-oriented scientific data integration framework consists of a set of interconnected service-enabled computation nodes. A service enabled computation node consists of both structured data repositories and service middleware. Structured data repositories, such as DBMS, GIS data stores, and data stream engines, store relational data, vector/raster geospatial data, and real-time data streams. They can be connected to physical devices to receive or pull streaming data. Structured data repositories are not necessarily independent; derived products from raw data can be generated and saved as additional data repositories. For example, the daily maximum/average ETo value of an hourly model output (in raster format) can be derived and saved in a GIS. Also monthly ETo average from weather stations can be implemented as either a regular view (named database query) or a materialized view (a physical instantiation of the query result) in a database. In addition, a stream engine (e.g., Ring Buffer Network Bus, RBNB [79]), whose primary functionality is to serve as a middleware for real-time data, can be used by certain client modules to update a database for archiving purposes. Middleware at the service level

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

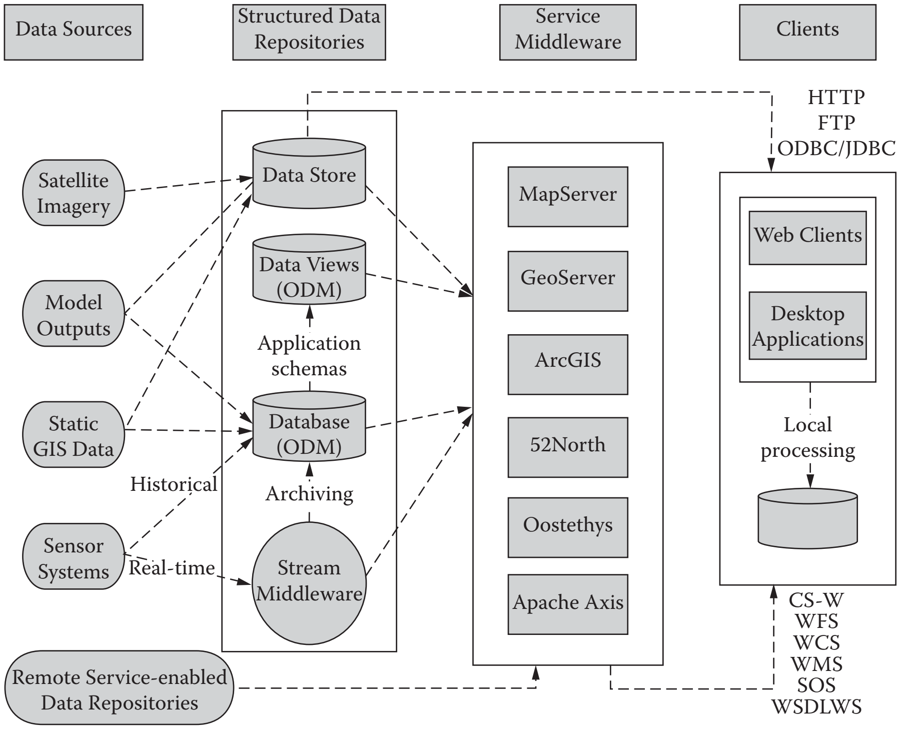

Figure 10.6 Conceptual architecture for geospatial data integration and interoperability

is responsible for extracting data from structured data repositories and provide it to clients in standard-compliant formats. Several commercial and open source service middleware systems, such as MapServer [81], GeoServer [29], THREDDS Data Server (TDS) [80], 52North [30], and so forth. are currently available. As such, structured data repositories should be formulated to work with the service middleware when possible. For example, PostgreSQL databases storing weather station measurement data can adopt the Community Observations Data Model (ODM) developed by CUAHSI [8] in formulating their table structures.

We note that the relationship between the structured data repositories, the service middleware, and their services are not necessary one-to-one. For example, both MapServer and GeoServer can connect to the PostgreSQL databases and provide WMS/WFS/WCS services. Similarly, TDS can provide WCS and OPeNDAP services from NetCDF data repositories. We also note that while Web service–based protocols are preferred for geospatial data integration in a Web application environment, more traditional communication protocols among the service-enabled computation nodes, such as distributed databases over TCP/IP, are not precluded. In addition, quite a few service middleware systems have the capability of connecting to remote data repositories either through database interfaces or service interfaces (WFS, WMS, WCS, SOS); thus they can integrate different data sources that are stored in local or remote data repositories and provide new services. While normally such type of integration is simple and limited in functionality, it could be appropriate in some applications or be used as parts of larger integration tasks. For example, MapServer can be configured to consume remote WFS services and use them along with local data in formulating the structure of a new WMS service.

In the next section, we will illustrate the incorporation of the various components of our scenario including the integration of real-time data.

10.4.3 Overall Integration and Results

As indicated at the beginning of this section, the goal in our scenario is the visualization, monitoring, and validation of model-based evapotranspiration for different eco-regions and selected locations in California. This goal is accomplished as follows. Using the integration tool, the user selects some weather stations to perform a comparison of ETo estimations against observed measurements at the given locations on an hourly basis (see Figure 10.7). The integration platform sends WFS requests to the relevant service endpoints to retrieve the station locations as well as the eco-region data, which are returned as vector features. It then sends a WCS request to the weather model output repository and retrieves the model output for the current time in NetCDF format. In general, the integration platform has to re-project the station geographical coordinates (usually given in latitude/longitude) to the model output projection (which is an equal-area based projection in the case of the WRF model [85]) before the corresponding station locations in the model output grids can be retrieved. The integration platform then loops through the time steps (hours in the figure) to retrieve

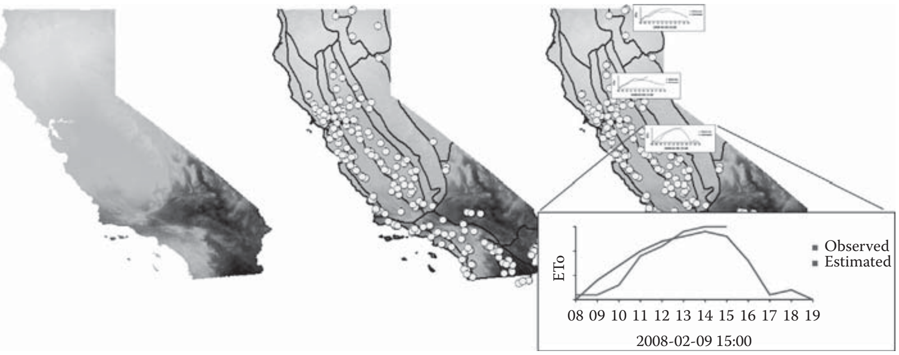

Figure 10.7 Main stages in the integration of sensor and model-generated data streams. Left: Estimated ETo map for the current hour; Center: Eco-regions and station locations overlayed; Right: Real-time charts for selected locations including the model prediction for a several-hour period and the actual observed values until current time.

```
◇ sml:SensorML
    ◇ sml:Sensor: id = CIMIS_S33_ETO
    ◇ name = CIMIS STATION 33 ETO
        ◇ Identification, ◇ Classification
        ◇ ReferenceFrame
        ◇ Input: name = eto
            ◇ Quantity: urn:ogc:def:phenomenon:eto
        ◇ Output: name = eto
            ◇ Quantity: urn:ogc:def:phenomenon:eto
```

Figure 10.8 Schematic SensorML document describing measured ETo

the ETo predictions at the station locations. The retrieved data can be rendered as charts by the integration platform for all desired stations for visualization purposes.

In real-time, the generated chart for each selected location also includes the observations from ground stations (from the CIMIS network [7] in our example). For this component, we use sensor web enablement (SWE)–related technologies. As already indicated in Section 10.2, both static and real-time sensor data can be provided through the Sensor Observation Service (SOS) interface; however, here we focus on representative sensor definitions and possible real-time access mechanisms. Following the SOS interface, a provider generates a capabilities document as shown in Figure 10.5. A DescribeSensor request for a particular detector produces a SensorML document describing the instrument that generates the stream. Figure 10.8 is a schematic depiction of part of a SensorML document for ETo measurements from a weather station system.

The response to a GetObservation request is an O&M document including a hyperlink that allows the client to open a connection to the data stream. An example of an O&M observation document is shown in Figure 10.9. A realization architecture would utilize a middleware data stream system as the entry point for all incoming stream data sources. This intermediate component would allow the implementation of various possible connections. An RBNB system [79] can be used as the entry point.

```
◇ om:Observation: id = CIMIS_S33_ETO
  ◇ Description: Observation with remote streaming result
  ◇ name = CIMIS STATION 33 ETO
  ◇ TimePeriod
    ◇ beginPosition: 2008-02-10T12:00:00:00
    ◇ endPosition: future
  ◇ Result
    xlink:href="http://comet.ucdavis.edu:9090/CIMIS/?ch=/S33/eto"
    xlink:role="application/octect-stream"
```

Figure 10.9 Example of an O&M observation response

In this case, the hyperlink shown in Figure 10.9 points directly to the corresponding channel in the RBNB server. However, various other types of connections are possible, including a TransducerML (TML) stream as a wrapper for the original, native stream, a TML stream as a wrapper for the RBNB stream, and the RBNB stream directly. The capabilities document would advertise the supported connection types so a client application can choose the one it is able to use.

In summary, as shown in Figure 10.7, right, both the prediction and the real-time ETo values can be displayed in chart form next to the respective station locations, and showed in the context of eco-regions (vector data) and current ETo maps (raster data), thus providing users and scientists with a vivid interface to monitor ETo values and easily compare them with weather model prediction over time.

It finally should be noted that the integration platform outlined above provides transparent and uniform access to heterogeneous geospatial data sources. However, in general, certain integration tasks such as resolving structural and semantic heterogeneities (see Section 10.2.4) still need to be explicitly realized at the client side and integration platform. These tasks include matching vector-based objects from two or more sources, selecting respective non-spatial attributes, and resolving general conflation aspects among data to be integrated. A viable approach to support such tasks is through scientific workflows (see Chapter 13), where the logic of conflict-resolving techniques is implemented in the form of actors.

10.5 Conclusions

With the amount of geospatial data growing at unprecedented rates, its effective sharing, exchange, and integration becomes a more critical necessity than ever before. We have seen that this goal involves not only dealing with various types of scientific data, but also integrating the increasing number of data and value-added services that are being deployed by geospatial communities in several important scientific application domains. The primary objective in this context is indeed a high degree of interoperability as a prerequisite for effective data integration and uniform and transparent data access.

In this chapter, we have reviewed emerging data integration requirements particularly in the context of the geosciences, where advancements in sensor and network technologies are placing an immense amount of diverse data at the scientist's disposal. We reviewed integration concepts from basic notions like data formats and metadata standards, to more comprehensive approaches including standards for interoperability and supporting Web-based technologies. We paid special attention to current efforts in the context of geospatial sensor data streams, amply exemplified with the enormous deal of data generated by air and space-borne instruments as well as numerous oceanic and ground sensor networks. With a practical environmental scenario, we illustrated an approach for integration and interoperability involving several of the components discussed in the chapter, which is in fact being developed in the context of the COMET Project [77]. Related projects, such as the Geoscience Network (GEON) [28] and the Science Environment for Ecological Knowledge (SEEK) [68], have also made great progress in building service-oriented architectures and portals that facilitate the efficient access to and integration of diverse geospatial datasets and repositories.

We have illustrated how current technologies, characterized by concerted efforts in standardization, are making the interoperability goal not only better defined but also effectively realizable in critical scientific application scenarios. Although much is still to be accomplished, especially in terms of the specification of ontologies in several areas of the geosciences, the science community can already take advantage of currently available infrastructures and technologies, and start benefiting from the progress underway.

Acknowledgements

This work is in part supported by the National Science Foundation under Awards No. IIS-0326517 and ATM-0619139.

References

[1] J. J. Bates. Exploratory climate analysis tools for environmental satellite and weather radar data (invited talk). In *Workshop on Management and Processing of Data Streams (MPDS 2003)*, 2003.

[2] C. Beeri, Y. Doytsher, Y. Kanza, E. Safra, and Y. Sagiv. Finding corresponding objects when integrating several geo-spatial datasets. In *Proc. 13th Int. Symp. on Geographic Information Systems*, 87–96, 2005.

[3] C. Beeri, Y. Kanza, E. Safra, and Y. Sagiv. Object fusion in geographic information systems. In *Proc. 30th Int. Conference on Very Large Data Bases*, 816–827, 2004.

[4] O. Boucelma, M. Essid, and Z. Lacroix. A WFS-based mediation system for GIS interoperability. In *Proc. 10th Int. Symp. on Advances in Geographic Information Systems*, 23–28, 2002.

[5] O. A. Bukhres and A. K. Elmagarmid. *Object-Oriented Multidatabase Systems*. Prentice Hall, 1996.

Please include publisher information

[6] C.-C. Chen, C. A. Knoblock, C. Shahabi. Automatically conflating road vector data with orthoimagery. *GeoInformatica*, 10(4):495–530, 2006.

To appear in Arie Shoshani and Doron Rotem (Editors): Scientific Data Management: Challenges, Existing Technology, and Deployment. ISBN 1420069802

[7] California Irrigation Management Information System (CIMIS). http://wwwcimis.water.ca.gov/cimis.

For all websites, please include date accessed if known.

[8] Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI). http://www.cuahsi.org/.

[9] G. da Rocha Barreto Pinto, S. P. J. Medeiros, J. M. de Souza, J. C. M. Strauch, and C. R. F. Marques. Spatial data integration in a collaborative design framework. *Commun. ACM*, 46(3):86–90, 2003.

[10] A. M. Ellison, L. J. Osterweil, L. Clarke, J. L. Hadley, A. Wise, E. Boose, D. R. Foster, A. Hanson, D. Jensen, P. Kuzeja, E. Riseman, and H. Schultz. Analytic webs support the synthesis of ecological datasets. *Ecology*, 87(6):1345–1358, 2006.

[11] A. K. Elmagarmid, M. Rusinkiewicz, and A. Sheth. *Management of Heterogeneous & Autonomous Database Systems*. Morgan Kaufman, 1999.

Please include location of publisher.

[12] Ecological Metadata Language (EML). http://knb.ecoinformatics.org/software/eml/.

Please check URL.

[13] Entrez. The GDB Human Genome Database. http://www.gdb.org/.

[14] T. Erl. *Service-Oriented Architecture (SOA): Concepts, Technology and Design*. Prentice Hall, 2005.

[15] Earth System Grid. http://www.earthsystemgrid.org/.

[16] ESRI. *ArcGIS 9*. ESRI Press.

[17] ESRI ArcInfo Product Description. http://www.esri.com/software/arcgis/arcinfo.

[18] ESRI Shapefile Technical Description, White Paper. Technical report, Environmental Systems Research Institute, Inc., 1998.

[19] M. Essid, F.-M. Colonna, O. Boucelma, and A. Bétari. Querying mediated geographic data sources. In *10th Int. Conference on Extending Database Technology*, LNCS 3896, Springer, 1176–1181, 2006.

[20] FGDC. Content Standard for Digital Geospatial Metadata (CSDGM), FGDC-STD-001-1998. http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata /base-metadata/, 2005.

Please check URL.

[21] Federal Geographic Data Commitee (FGDC). http://www.fgdc.gov, 2007.

[22] I. Foster and C. Kesselman. *The Grid 2: Blueprint for a New Computing Infrastructure (2nd Edition)*. Morgan Kaufmann, 2004.

[23] I. Foster. Service-oriented science: Scaling e-science impact. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, 9–10, 2006.

Please include location of publisher.

[24] Global Biodiversity Information Facility. http://www.gbif.org/.

[25] GDAL — Geospatial Data Abstraction Library. http://www.gdal.org/.

[26] The GDB Human Genome Database. http://www.gdb.org/.

[27] Global Earth Observation Grid. http://www.geogrid.org/.

[28] Geosciences Network (GEON). http://www.geongrid.org/.

[29] GeoServer. http://geoserver.org.

[30] Geospatial Open Source Software GmbH. http://52north.org/.

[31] E. P. Glenn, A. R. Huete, P. L. Nagler, K. K. Hirschboeck, and P. Brown. Integrating remote sensing and ground methods to estimate evapotranspiration. *Crit. Rev. in Plant Sciences*, 26(3):139–168, 2007.

[32] Global Earth Observation System of Systems. http://www.epa.gov/geoss/.

[33] Geographic Resources Analysis Support System (GRASS). http://grass.osgeo.org/.

[34] J. Gray and A. Szalay. The world-wide telescope, an archetype for online science. Technical report msr-tr-2002-75, Microsoft Research, 2002.

[35] GriPhyN. Grid Physics Network. http://www.griphyn.org/. **Please check URL.**

[36] A. Y. Halevy, N. Ashish, D. Bitton, M. J. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka. Enterprise information integration: successes, challenges and controversies. In *Proceedings of the ACM SIGMOD Int. Conference on Management of Data*, 778–787, 2005.

[37] Hierachical Data Format (HDF5). http://hdf.ncsa.uiuc.edu/products/hdf5/.

[38] IBM. *DB2Spatial Extender and Geodetic Data Management Feature User's Guide and Reference*. IBM Corp., 2006.

[39] J. C. Iliffe. *Datums and Map Projections: For Remote Sensing, GIS and Surveying*. Whittles Publishing, 2000.

[40] M. B. Jones, M. P. Schildhauer, O. Reichman, and S. Bowers. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology Evolution and Systematics*, 37:519–544, 2006.

[41] J. L. S. Jr., A. A. Taylor, P. R. Hershberg, and R. Bunge. Disseminating national weather service digital forecasts using open geospatial standards. In *Proceedings of the 23rd AMS Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, page 3B.9, AMS Press, 2007. **Au: need author last name.**

[42] R. Lake, D. Burggraf, M. Trninic, and L. Rae. *Geography Mark-Up Language: Foundation for the Geo-Web*. Wiley, 2004.

[43] Y. Lassoued, M. Essid, O. Boucelma, and M. Quafafou. Quality-driven mediation for geographic data. In *Proceedings of the Fifth International Workshop on Quality in Databases, QDB 2007*, 27–38, 2007.

[44] J. Lee, Y. Lee, S. Shah, and J. Geller. HIS-KCWATER: context-aware geospatial data and service integration. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC)*, 24–29, 2007.

[45] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind. *Geographic Information Systems and Science*. Wiley, 2005.

Please include location of publisher.

[46] C. Murray. *Oracle Spatial Developer's Guide, 11g Release 1*, 2007.

[47] Semantic Web for Earth and Environmental Terminology (SWEET). http://sweet.jpl.nasa.gov/, 2005.

[48] National Ecological Observation Network (NEON). http://www.neoninc.org/.

[49] Network Common Data Form (NetCDF). http://www.unidata.ucar.edu/software/netcdf/.

[50] GeoTIFF. http://www.remotesensing.org/geotiff/.

[51] M. Neteler and H. Mitasova. *Open Source GIS: A GRASS GIS Approach*. Springer, 2007.

Please include location of publisher.

[52] National Spatial Data Infrastructure (NSDI). http://www.fgdc.gov/nsdi/nsdi.html.

[53] IEEE standard computer dictionary: A compilation of IEEE standard computer glossaries., 1990.

[54] OGC Interoperability Institute. http://www.ogcii.org/.

[55] OGC. OpenGISR® Implementation Specification for Geographic information — Simple feature access — Part 1: Common architecture. version 1.2.0. http://www.opengeospatial.org/standards/sfa, 2006.

[56] OGC. OpenGISR® Geography Markup Language (GML) Encoding Specification. version 3.1.1. http://www.opengeospatial.org/standards/gml, 2007.

[57] OGC. OpenGISR® Sensor Web Enablement: Architecture Document. http://www.opengeospatial.org/pt/14140, 2007.

[58] OGC. Catalogue Service Implementation Specification. http://www.opengeospatial.org/standards/cat.

[59] OPeNDAP: Open-source Project for a Network Data Access Protocol. http://www.opendap.org/.

[60] T. Ormsby, E. Napoleon, and R. Burke. *Getting to Know ArcGIS Desktop: The Basics of ArcView, ArcEditor, and ArcInfo*. Esri Press, 2004.

Please include location of publisher.

[61] G. Percivall and C. Reed. OGC Sensor Web Enablement Standards. *Sensors & Transducers Journal*, 71(9):698–706, September 2006.

[62] PostGIS. http://postgis.refractions.net/.

[63] PROJ.4 - Cartographic Projections Library. http://proj.maptools.org/.

[64] P. Rigaux, M. Scholl, and A. Voisard. *Spatial Databases with Application to GIS*. Morgan Kaufmann, 2002.

[65] G. K. Rutledgea, J. Alpert, and W. Ebisuzaki. Nomads: A climate and weather model archive at the national oceanic and atmospheric administration. *Bulletin of the Am. Meteorological Society*, 87(3):327–341, 2006.

[66] E. Safra, Y. Kanza, Y. Sagiv, and Y. Doytsher. Efficient integration of road maps. In *14th Int. Symp. on Geographic Information Systems*, 59–66, 2006.

[67] Sloan Digial Sky Survey (SDSS). http://www.sdss.org/.

[68] Science Environment for Ecological Knowledge (SEEK). http://seek.ecoinformatics.org/.

[69] V. Sehgal, L. Getoor, and P. Viechnicki. Entity resolution in geospatial data integration. In *14th Int. Symp. on Geographic Information Systems*, 83–90, 2006.

[70] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, June 2002.

Please include location of publisher.

[71] A. P. Sheth. Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In *Interoperating Geographic Information Systems*, 5–30. Kluwer, 1999.

[72] A. P. Sheth and J. A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[73] W. Shi, P. Fisher, and M. F. Goodchild. *Spatial Data Quality*. CRC, 2002.

Please include location of publisher.

[74] J. E. Stern. State Plan Coordinate System of 1983. NOAA Manual NOS NGS 5, March 1990.

[75] K. Stolze. SQL/MM Spatial — The Standard to Manage Spatial Data in a Relational Database System. In *Tagungsband der 10. BTW-Konferenz*, LNI 26, 247–264, 2003.

[76] A. Szalay and J. Gray. 2020 computing: Science in an exponential world. *Nature*, (440):413–414, 2006.

[77] The COMET Project, COast-to-Mountain Environmental Transect. http://comet.cs.ucdavis.edu.

[78] The Open Geospatial Consortium (OGC). http://www.opengeospatial.org.

[79] S. Tilak, P. Hubbard, M. Miller, and T. Fountain. The ring buffer network bus (RBNB) dataturbine streaming data middleware for environmental observing systems. In *Proc. 3rd Int. Conf. on e-Science and Grid Computing (e-Science 2007)*, 125–133, IEEE, 2007.

[80] University Corporation for Atmospheric Research. Thematic Realtime Environmental Distributed Data Services. http://www.unidata.ucar.edu/projects/THREDDS/.

[81] University of Minnesota. Mapserver, http://mapserver.gis.umn.edu.

[82] U.S. Census Bureau. Topologically Integrated Geographic Encoding and Referencing system (TIGER). http://www.census.gov/geo/www/tiger/.

[83] J. M. Ware and C. B. Jones. Matching and aligning features in overlayed coverages. In *Proc. 6th Int. Symp. on Advances in Geographic Information Systems*, 28–33, 1998.

[84] A. Woolf, K. Haines, and C. Liu. A web service model for climate data access on the grid. *Int. J. High Perform. Comput. Appl.*, 17(3):281–295, 2003.

[85] Weather Research & Forecasting Model (WRF). http://www.wrf-model.org

[86] Q. Yang, J. Snyder, and W. Toble. *Map Projection Transformation: Principles and Applications*. CRC, 1999.

[87] P. Ziegler and K. R. Dittrich. Three decades of data integration — all problems solved? In *Building the Information Society, IFIP 18th World Computer Congress, Topical Sessions*, 3–12, 2004.