49
Metadata and data catalogues

S C GUPTILL

The use of GIS technologies for analysing spatial problems has become pervasive and, in the process, has created a demand for vast amounts of digital geospatial data. Two coupled developments – the use of metadata and the provision of data catalogues – are helping to create an environment to accommodate today's requirements for geospatial data. Metadata provide descriptive information about the producer, content, quality, condition, and other characteristics of a given item. Agreements on common core metadata elements will hasten the development of efficient software tools to lower the present high cost of metadata entry. Collections of metadata records are the basis for data catalogues and clearinghouses that can be searched by potential data users. Rapid advances in Internet technology are making such data clearinghouses easy to use for data discovery. But we still await an effective set of tools that will allow users to determine the fitness for use characteristics of any given dataset.

1 INTRODUCTION

Geospatial data, or data tied to a location on the Earth, are critical to solving today's complex environmental, economic, and social problems. The use of GIS technologies for analysing spatial problems has become pervasive and, in the process, has created a demand for vast amounts of digital geospatial data. To meet that demand, the data that describe the characteristics of geographical space must be current and of known quality. The data must be easily accessible and able to be integrated, generalised, and manipulated with understood or predictable results. Meeting all of these requirements is difficult today. Two coupled developments – the use of metadata and the provision of data catalogues – are helping to create an environment to accommodate today's requirements for geospatial data.

As technology provides faster and more efficient ways to transmit and process geographical data, data producers and users are realising the potential value of exchanging and sharing spatial data. The ability to use existing spatial data is important to organisations which are trying both to share data and to increase benefits from their data collections. This ability depends not only on being able to find data, but also on understanding the characteristics and quality of the data. Metadata, or data about data, provide information such as the characteristics of a dataset, its history, and the organisations to contact to obtain a copy of it (see also Goodchild and Longley, Chapter 40). Standardised metadata elements provide a means to document datasets within an organisation, to contribute to catalogues of data that help individuals find and use existing data, and to help users understand the contents of datasets that they receive from others.

A collection of metadata records combined with data management and search tools forms a data catalogue. The architecture of these data catalogues can vary from tightly held internal data indexes to widely accessible distributed catalogues spread across the Internet. The catalogue function can go far beyond the basic listing of content. Extensions that allow for direct ordering or download of data, user comment or feedback, as well as requests for new data collections, are all possible. Metadata and data catalogues are thus the means by which data users can more economically find, share, and use geospatial data.

2 METADATA

Metadata provide descriptive information about the producer, content, quality, condition, and other characteristics of a given item. These items may be as varied as reports, books, maps, photographs, and digital datasets. Metadata elements about each of these items would probably have some common terms. All might have an author or producer and a date of publication. However, each might have unique elements. The book might have a unique ISBN number, the map a scale factor, the photograph an exposure setting, and the digital dataset a format description. The creators of a metadata specification (the list of descriptive terms to be used to describe the set of items under consideration) face the difficult task of not only categorising the set of items to be covered by the metadata specification, but also of describing each element in it. The choice of elements will depend on the intended use, or purpose, of the metadata. In addition, the metadata must complement other descriptive systems in use (for example, the catalogue systems of libraries).

2.1 Purposes of metadata

Metadata descriptions endeavour to provide a robust list of common terminology and definitions for data elements. Typically, these elements have been selected to provide information about one of these four topics:

- availability – data needed to determine the sets of data that exist for a given subject or geographical location;
- fitness for use – data needed to determine if a set of data meets a specified need;
- access – data needed to acquire an identified set of data;
- transfer – data needed to process and use a set of data.

When elements addressing all four of these topics exist, they form a continuum where a user cascades through a pyramid of choices when determining which data are available, evaluating the fitness of the data for a particular use then accessing, transferring, and receiving the data. However, many metadata specifications address only a subset of these topics. The vast majority of information in existing data catalogues focuses on availability, with a minimal amount of information on access and transfer. Very little information on fitness for use is available. This places the user in the unenviable position of having to acquire and analyse the data to determine if it fits their specific needs.

Metadata can be used in at least three different contexts: catalogues, management records, and accompanying a dataset. Each context may require a different form of metadata in order to facilitate its use. Information appropriate for use in one context may be less useful in another (for example 'price' is useful in a catalogue but less useful when accompanying a dataset that has already been purchased). The fact that different metadata applications require different sets of metadata elements makes any standardisation process more complex.

2.2 Data versus metadata

What is the difference between data and metadata; between metadata and meta-metadata? In many ways any differences are in the eye of the beholder (Lillywhite 1991). There are no firm rules upon which to base a decision. Data are generally viewed as elements that model or represent real-world phenomena, for example a line segment that represents the location of a road. As collections of these elements are put together (e.g. the national highway system), so information about their collection (i.e. metadata) comes into play. As accretion of data occurs (by adding new data themes or areas of coverage for example), so the need for more metadata grows as well.

Consider an inventory of digital map data (based on individual map sheets) covering an entire country. At a minimum, there are: metadata about the entire collection; metadata about each map sheet; and metadata about each theme (roads, rivers, elevation, etc.) within each sheet. GIS users could argue that they need metadata about each feature within each theme, particularly if such features differ, say, in accuracy or date of collection. It is apparent that the amount of metadata could approach or exceed the amount of actual data. It is also apparent that the collection and management of metadata is a major undertaking in itself.

2.3 Scope of a metadata specification

The first major decision to be made by those implementing a metadata specification is to determine the scope of the information that will be included within its descriptive parameters. For

example, should a specification apply only to digital data or to non-digital data as well? Should the metadata include data that are completed, in progress, or planned? Adding elements for non-digital forms would complicate the effort. However, many users are interested in obtaining data regardless of the medium on which they are stored. A standard is therefore needed to provide uniform descriptions of data on any media. By definition, geospatial data can be presumed to have some form of spatial referencing as a component. But there are many possible means of describing spatial location (see Seeger, Chapter 30). These means include georeferenced coordinates (such as longitude–latitude coordinates), coordinates whose relationship to the Earth is not known (such as some survey data or remotely-sensed data that have not been corrected geometrically), other means of describing locations (such as street addresses and mile markers), and indirect positional references through objects that have a known location (such as references to counties or wells).

In addition, ancillary datasets may have information important to spatial analyses and may be relatable to spatial data but may not include any spatial references. For example, a lexicon of soil classification may contain detailed information about the soil characteristics and appropriate uses, but contain no spatial information. A soil map, where areas of homogeneous soils are delimited, would need to reference this more detailed set of descriptive information about each soil unit.

The scope of the metadata specification, the detailed elements of that specification, the data resulting from collecting the information, and the software used to search the resulting database all interact, sometimes in ways not anticipated by the designers. Making the spatial referencing schemes more inclusive can have (perhaps unintended) effects on the searching of spatial data clearinghouses. Consider this search, via the World Wide Web (WWW), of the Global Environmental Network for Information Exchange in the United Kingdom (the GENIE Service; *http://www-genie.mrrl.lut.ac.uk*). The search form allows the user to specify 'Places'. The database returns metadata entries which the metadata provider has referenced to a certain geographical area. A query with the following criteria – Places: Isle of Wight – returned three entries as given below:

British Geological Survey – Hampshire Basin Tertiary Sands data
Economic and Social Research Council – Electoral Dynamics Files
Economic and Social Research Council – Wessex Survey of Marathon Runners

The first two entries – studies of sand composition and election results, respectively – for the Isle of Wight, are not surprising. However, the third (a survey on the habits and mind sets of people that run marathons) gets included because a marathon race was conducted on the Isle of Wight, a fact peripherally related to the topic of the report. The further interactions of metadata specifications, data catalogues, and search engines will be examined in a later section.

2.4 Content

Once the scope and purpose of the metadata specification has been generally agreed, then specific content supporting those goals can be specified. In the USA, the Federal Geographic Data Committee (FGDC) went through this process and developed a content specification for geospatial metadata (FGDC 1994). The FGDC metadata standard allows documentation of datasets that locate information by coordinates, by addresses and, indirectly, by using features of a known location, such as a county. The standard can be used to document ancillary datasets that can be linked to spatial datasets, individual datasets, or a series of datasets. It allows documentation of data that are completed, in progress, or planned. The major data element categories are listed in Table 1.

These metadata categories may be compared with those developed by the Australians for the FINDAR (Facility for Interrogating the National Directory of Australian Resources) directory system. The FINDAR directory was developed by the National Resource Information Centre to 'provide flexibility in describing datasets that exist in a wide range of

Table 1 FGDC categories of spatial metadata.

Identification information
Spatial reference
Status information
Metadata reference information
Source information
Processing history information
Distribution information
Entity/attribute information
Contact information

formats (numeric databases, bibliographies, indexes, maps, etc.) and in meeting the needs of different types of users' (Johnson et al 1991). The FINDAR list of major metadata descriptors is given in Table 2.

Table 2 FINDAR metadata categories.

Identification
Data items
Spatial identification
Spatial coverage
Dataset information
Data currency
Data lineage and quality
Ordering information
Keywords
Organisation/position information

At this general level, these lists are quite similar. Indeed the similarity holds even with comparisons to metadata specifications from non-spatially orientated organisations such as nuclear weapons laboratories (Lownsbery and Newton 1996) or molecular science (Keller and Jones 1996). Since, without formal effort, we have almost arrived at a consensus of opinion on metadata content, is it possible to agree on a formal standard? Efforts on standardising metadata are discussed in section 2.5 and, as is often the case, 'the devil is in the details'.

2.4.1 Choice of data elements

Metadata enable data users to identify existing data, identify the appropriate dataset for a particular use, find information to access datasets, and find information to transfer datasets for processing and use. Although the potential benefits of metadata are obvious, several impediments hinder uniform adoption and use. These centre on two topics: agreement as to which metadata elements to collect; and the level of effort required to collect whatever metadata are required. Obviously the two items are linked. From the user perspective, detailed metadata are highly desirable. More information will help the user to determine the suitability of a given dataset. However, from the data producer's perspective, creating and maintaining metadata costs money (sometimes large amounts of it); metadata elements should thus be constrained to a cost-effective set. In addition, significantly more resources may be required to collect metadata about 'legacy' data holdings than datasets collected using techniques that capture the metadata as a part of the overall data collection process.

Choosing the elements that make up a metadata standard involves balancing several factors. The standard should be complete and support cataloguing and transfer activities; it also should be simple and provide only the minimum elements necessary. A large, complex standard is more burdensome, is less likely to be used, and presents more difficulties for verifying compliance and enforcement. However, increased metadata elements provide additional information and thus help a user to determine whether the data being described are appropriate for use. One way to deal with this problem is to declare some elements to be mandatory (or core), and others as optional. This strategy is less than ideal, as the producers of metadata may only supply information for the mandatory fields, leaving the optional ones blank. If the metadata item is essential for satisfying one of the four major purposes of the metadata, then it should be made mandatory.

2.4.2 Collecting metadata

As noted in section 2.2, there are no set rules on which data should be characterised by metadata descriptions. Certainly datasets (i.e. some logical collection of information based on content or geographical extent) should be described by metadata. But what about the components of the dataset? Each theme of a digital topographic map (such as roads, hydrography, elevation, and boundaries) could warrant a separate metadata description. It is possible to imagine metadata that could be associated with each feature in a dataset or even with attributes or attribute values of a feature. Clearly there is a practical limit to this cycle of data description. That limit is often reached by the effort required to collect the information.

Many of the metadata that are collected today must be entered manually, item by item. This process requires not only personnel with the appropriate information resources needed to place data in the metadata fields (such as written records describing the dataset, appropriate collection procedures, and availability of suitable formatting methods), but also tools to ensure that all the appropriate elements for a given dataset are filled. Software tools that ease this latter task have been developed. For example, the Biological Resources Division of the United States

Geological Survey (USGS) has developed a software package called 'MetaMaker'. This is a metadata entry program that uses Microsoft Access database tables to collect and organise metadata information. The user is provided with data entry forms that represent the various sections and compound elements of the FGDC metadata standard. The National Aeronautical and Space Administration (NASA) has developed a similar set of tools to collect metadata in its Directory Interchange Format (DIF). Given the complex set of dependencies that can exist in a set of metadata standards (for example, a contact person must have a phone number, or a conic map projection must have standard parallels), such software is invaluable in the data entry process. The University of Wisconsin maintains a WWW site that has linkages to a large number of software tools for metadata collection (*http://badger.state.wi.us/agencies/wlib/sco/pages/metadata/metadata.html*). However, significant resources are still needed in order to document properly large databases which have little or no metadata associated with them. The processing of legacy datasets is a major concern of many organisations because of the resources needed to accomplish this task.

A better solution henceforth is to incorporate metadata collection as part of the entire data collection process. In this scenario, the data collection software would have embedded components that collect the information needed for the various metadata fields. In some cases, the information would still need to be entered by an operator (such as descriptive text about the dataset). However other information, such as coordinate system, file format, or lineage, could be automatically calculated or stored by the software and used to populate the metadata fields. Efforts in this regard have focused on user-created macro routines as 'add-ons' to GIS packages. The GIS vendors have yet to incorporate comprehensive metadata gathering tools (see Sondheim et al, Chapter 24; Veregin, Chapter 12). Nor might one expect them to do this until a standard set of metadata is in widespread use.

2.5 Standardisation

Agreement on standardised terminology and definitions would provide benefits not only in the automation of the collection of metadata, but would also be exceedingly valuable in searching data catalogues. Several metadata 'standards' are in widespread use around the world. They will be briefly described here, followed by a discussion of efforts to consolidate the various activities.

2.5.1 MARC, Dublin Core, and Warwick Framework

Metadata classification and collection activities most closely resemble the tasks performed by librarians as they classify their holdings and create catalogue entries for all the items in a library collection (see Adler and Larsgaard, Chapter 64). Since the 1960s, librarians have been using computers to create electronic catalogues. Under the auspices of the National Information Standards Organisation (NISO), an American National Standards Institute (ANSI)-accredited organisation that develops standards specifically for the library, information services, and publishing sectors, a set of standards known as USMARC has evolved. The machine-readable cataloging (MARC) activities are international in scope, with various countries adopting country-specific variations, leading to standards such as UKMARC and FINMARC.

The USMARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. A USMARC record involves three elements: the record structure, the content designation, and the data content of the record. The structure of USMARC records is an implementation of national and international standards: for example, Information Interchange Format (ANSI Z39.2 1994) and Format for Information Exchange (ISO 2709 1996). The content designation, the codes and conventions established to identify explicitly and further characterise the data elements within a record and to support the manipulation of those data, is defined in the USMARC formats. The content of most data elements is defined by standards outside the formats (e.g. Anglo-American Cataloguing Rules, Library of Congress Subject Headings, National Library of Medicine Classification). The content of other data elements, coded data for example, is defined in the USMARC formats. A USMARC format is a set of codes and content designators defined for encoding machine-readable records. Formats are defined for five types of data: bibliographic, holdings, authority, classification, and community information. The USMARC formats are communication formats, primarily designed to provide specifications for the exchange of bibliographic and related information

between systems. They are widely used in a variety of exchange and processing environments. As communication formats, they do not require internal storage or display formats which are specific to individual systems. The Library of Congress maintains the USMARC formats in consultation with various user communities (Library of Congress 1996).

The number of elements, codes, and element relationships are large and complex within the USMARC system. In addition, USMARC was largely designed to catalogue physical objects (such as books, maps, records, movies, or pictures) and not electronic datasets. To address these and other issues, a metadata workshop sponsored by the Online Computer Library Centre (OCLC) and the National Centre for Supercomputing Applications (NCSA) convened 52 selected researchers and professionals from librarianship, computer science, text encoding, and related areas in March 1995, in order to advance the state of the art in the development of resource description (or metadata) records for networked electronic information objects. The primary deliverable from the workshop was a simple resource description record composed of a set of 13 metadata elements, named the Dublin Metadata Core Element Set (or Dublin Core, for short) by the workshop participants. The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet (Weibel et al 1995). The 13 elements are listed in Table 3.

A subsequent meeting in Warwick, England, in April 1996 dealt with the development of a formal syntax and deployment strategy for the Dublin Core. This led to the creation of the Warwick Framework.

Table 3 The Dublin Core.

Subject
Title
Author
Publisher
Other agent
Date
Object type
Form
Identifier
Relation
Source
Language
Coverage

This is a container architecture for bringing together different packages of metadata that are separately accessible and maintainable. It allows for deployment of the Dublin Core as well as for extensibility for other types of metadata not part of the Dublin Core (Library of Congress 1997).

2.5.2 Directory Interchange Format (DIF)

The DIF is a *de facto* standard developed by the satellite remote-sensing community which is used to create directory entries to describe a group of data. It has been widely used by NASA and NOAA (National Oceanic and Atmospheric Agency) to create catalogues of their large holdings of satellite information and global change datasets. The required elements of DIF are as shown in Table 4. Twenty-seven optional elements are used to describe the characteristics of satellite data. In addition, seven more fields were added to DIF to facilitate compliance with the FGDC metadata standard (NASA 1996).

Table 4 DIF required elements.

Directory entry identifier
Directory entry title
Parameters
Originating centre
Data centre (name, dataset ID, contact person)
Summary

2.5.3 Government Information Locator Service (GILS)

The United States Office of Management and Budget (OMB) is promoting the establishment of an agency-based GILS to help the public locate and access information throughout the Federal government. As part of the Federal role in the National Information Infrastructure, GILS will identify and describe information resources throughout the Federal government and provide assistance in obtaining the information. It will be decentralised and will supplement other agency and commercial information dissemination mechanisms. The public will use GILS directly or through intermediaries, such as the Government Printing Office (GPO), the National Technical Information Service, the Federal depository libraries, other public libraries, and private sector information services. Direct users will have access to a GILS Core accessible on the Internet without charge. Intermediate access may include kiosks, free telephone numbers, electronic mail, bulletin boards,

facsimile, and off line media such as floppy disks, CD-ROM, and printed works (Christian 1996).

GILS will use standard network technology and the ANSI Z39.50 standard (the same standard used by USMARC) for information search and retrieval so that information can be retrieved in a variety of ways. Direct users will eventually have access to many other Federal and non-Federal information resources, linkages to data systems, and electronic delivery of information products.

One of the GILS components is a set of locator records that reside on GILS-accessible servers and are further identified by agencies as belonging to the GILS Core. These are required to be maintained by Federal agencies having significant information holdings, where each record describes part of the agency holdings. These core locator records will be accessible comprehensively in the GPO Access system but can also be aggregated by direct users of GILS to provide selective views of Federal government holdings. The list of GILS Core metadata elements is available at the GILS WWW site (*http://www.usgs.gov/gils/prof_v2.html*).

2.5.4 Federal Geographic Data Committee (FGDC)

In June 1994, the FGDC adopted its Content Standards for Digital Geospatial Metadata (FGDC 1994). This standard was the end result of a two-year process of design, drafting, and public review. The standard was the first focused effort on specifying the information content of metadata for a set of geospatial data. The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data, to determine the fitness of a set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data, and to transfer successfully the set of geospatial data. As such, the standard establishes the names of about 220 data elements and compound elements to be used for these purposes, the definitions of these data elements and compound elements, and information about the values that are to be provided for the data elements. The standard does not specify the means by which this information is organised in a computer system or in a data transfer, nor the means by which this information is transmitted, communicated, or presented to the user.

The original FGDC standard contains a large number of mandatory elements. This has caused concern among, primarily, data producers on the amount of resources that will be needed to collect the required metadata. In the revised standard (to be completed in October 1998), elements that are identified as 'mandatory' are required to have values provided by the metadata provider to describe completely the geospatial dataset. Values for 'optional' elements may be provided at the discretion of the metadata implementer. Elements identified in the original standard as 'mandatory if applicable' are identified as 'mandatory' in this proposed revision. The term 'depends on use' indicates that the element could be mandatory or optional depending on how it is used in describing a dataset. Disposition refers to the importance and scope of an element's applicability, which is identified through the use of the terms 'core', 'standard', and 'user extensions'. Core elements are of critical importance to maintaining a comprehensive and consistent clearinghouse of metadata. They are generic enough in meaning and intent to describe any geospatial dataset regardless of any uniqueness that may exist because of differences in discipline, subject matter, or dimensionality. In this document, the elements identified as core are so noted in the proposed revision of the standard. Standard elements are those to be used to provide additional information about a geospatial dataset, but their population in the clearinghouse is not essential. Only core and standard elements are identified in this revised standard; no elements identified as a user extension currently appear in this proposed revision of the standard. The specific community that advocates a given enhancement should propose user extensions for inclusion in the standard. The proposed revised FGDC Content Standard for Geospatial Metadata can be found at the FGDC WWW site (*http://www.fgdc.gov/metadata/metadata.html*).

2.6 Interoperability among standards

The above set of examples shows a fair degree of content overlap. Each standard arose from a given constituent group, achieved consensus support, and has amassed a fair amount of metadata in compliance with the standard. However, there are now efforts to try to consolidate these disparate activities. While some of these efforts are directed towards creating a new, all encompassing, standard (see section 2.5.3), others are focused on working with the existing standards and making them interoperable (for a general discussion, see Sondheim et al, Chapter 24). One method is to

create 'crosswalks' between the various standards, that is identifying analogous or closely related items in each standard and using the same data value in each of those fields. This allows a user query to be 'translated' to be applicable across a number of metadata standards. The Environmental Information Services WWW site (*http://www.esdim.noaa.gov/fgdcdif.html*) shows a crosswalk between the FGDC and DIF standards that was created by the NOAA (Barton 1996). Translations such as this one allow the NOAA Data Catalog to be accessed and viewed utilising a variety of standard metadata formats. The user may view the data descriptions in several formats including DIF, FGDC Metadata, GILS, and a generic format. The data descriptions are stored in the DIF format and a format mapping application converts them to the user specified format at display time. NOAA provides the data descriptions in various formats to satisfy the requirements of several United States federal government programmes (Barton 1996).

A significant amount of resources is needed in order to support the development and maintenance of writing pair-wise data interface programs. As with any data interface approach, the number of interfaces required (and the effort required) increases as the square of the number of communicating systems. A different approach, data mediation, is being explored by a number of researchers (Renner et al 1996). The data mediator is a computer program that translates data between two systems with different data schemas. The difference between a mediator and the translation programs currently in use is that the mediator automatically generates data translations from descriptions of the data in the source and receiver systems. In this approach, metadata catalogue developers describe their data schema once. The description can then be used by the mediator to communicate with as many other systems as desired. For this method to work, all the data descriptions must be written in a formal language, written using a common vocabulary, and be adequate for the level of translation desired.

2.7 Standardising the standards

A great deal of activity is centred upon formalising the concepts expressed in the Dublin Core. This is being undertaken by a range of computer science, text mark-up, and library science professionals. The Internet Engineering Task Force (IETF) is the organisational home for the work on the Dublin Core. An IETF Working Group on metadata is seen as an effective way of involving the members of the computer and networking communities that must implement and use the standards (Weibel et al 1995). A number of different development directions are being explored, including:

- expansion of the Dublin Core to include other object types;
- expansion of the Dublin Core to include more functionality, other than resource discovery, such as archive control;
- establishing standardised methods for extensibility;
- testing and refining existing Core elements;
- promoting semantic interoperability across disciplines and languages;
- defining mechanisms to support richer description and linkages to other description models.

One ongoing research project is to test the provision of Dublin Core metadata in the header records of HTML (hypertext mark-up language) documents (Hakala et al 1996). This approach would allow for efficient indexing of electronic documents. Once a recommendation on how to use HTML tags to provide Dublin Core metadata is completed, simple templates for editor applications might be created in order to reduce the effort required for metadata collection.

The International Standards Organisation (ISO) is a worldwide federation of national standards bodies from approximately 100 countries (see Salgé, Chapter 50). The standards development work of the ISO is carried out through Technical Committees (TC). ISO TC 211 Geographic Information/Geomatics (*http://www.statkart.no/ isotc211*) is the committee working on standardisation in the field of digital geographical information. One of the projects of TC211 deals with metadata. Project 15046-15 'Geographic information – metadata' has as its scope the 'definition of the schema required for describing geographic information and services'. This document will provide a metadata content standard like the FGDC standard.

This standard has been based upon the work carried out by a number of countries which have created their own metadata standards. These include: the Australia and New Zealand Land Information Council (ANZLIC) Working Group on

Metadata (Core Metadata Elements available at *http://www.anzlic.org.au*); the Canadian Directory Information Describing Digital Georeferenced DataSets; the European Committee for Standardisation (CEN) Standard for 'Geographic Information – Metadata' (available at *http://forum.afnor.fr/afnor/WORK/AFNOR/GPN2/Z13C/PUBLIC/WEB/ENGLISH/pren.htm*); and the US FGDC 'Content Standard for Geospatial Metadata'. It is proposed that in the ISO Metadata Standard the elements of the metadata standard should be organised around a minimum set of 'core' elements, a larger list of standardised 'optional' elements, and a set of rules that define the name, description, type, and domain for user-defined elements (used to document metadata items not provided for in the formal standard). The core elements are either mandatory (shall be present) or conditional (shall be present if the dataset exhibits the characteristics defined by that element). The major categories of metadata in the proposed standard are shown in Table 5 (ISO 1997). The draft of ISO Standard 15046-15 is proceeding through the ISO review and approval process with a scheduled completion date in early 1998 (Danko 1997).

However ISO TC211 is not the only ISO body concerned with metadata (see also Salgé, Chapter 50). ISO/IEC Joint Technical Committee 1, Subcommittee 14 (Data Element Principles) has developed a six part standard addressing the standardisation of metadata (Newton 1996). The 'Specification and Standardisation of Data Elements – ISO 11179' describes rules, principles, and guidelines for classifying, attributing, describing, naming, and registering data elements. A set of standardised attributes, some with sample values, prescribes the base set of information to be recorded about each data element. In addition, procedures for registering standard data elements provide a mechanism for sharing this standardised information among organisations.

Since ISO 11179 is an approved standard, the efforts of ISO TC211 will need to be harmonised with it. However, the activities of the IETF group seem to be independent of the ISO activities. Given the large number of groups concerned with metadata, it is difficult to gain an integrated view of what is happening in the field. The International Federation of Library Associations and Institutions (IFLA), The Hague, Netherlands, maintains a comprehensive WWW site (*http://www.nlc-bnc.ca/ifla/III/metadata.htm*) devoted to the various standards and activities in the metadata field. The community would benefit if these standard activities, as well as any other related activities, were considered at a single venue.

Table 5 Major categories of the proposed ISO TC211 Metadata Standard.

Identification	Mandatory
Data quality	Conditional
Spatial data representation information	Mandatory
Spatial reference	Conditional
Feature and feature attribute	Conditional
Distribution	Conditional
Metadata reference	Mandatory

3 DATA CLEARINGHOUSE

Collections of metadata records (regardless of form) are the basis for data catalogues that can be searched by potential data users. Establishing a data catalogue can be a major project, and requires the cataloguing organisation to:

- identify those programmes or individuals that produce, manage, or disseminate data worthy of inclusion in the catalogue;
- identify what geospatial data an organisation holds, where its data are located, and in what condition they are held;
- identify priorities for documenting previously collected data based on demand, geographical coverage, uniqueness, or other criteria to be determined by the organisation;
- create the metadata for new geospatial data;
- ensure that the metadata comply with applicable standards, to the extent possible for the organisation;
- provide training and technical assistance to metadata producers, quality assurance personnel, and end-users;
- update and maintain the metadata as necessary;
- specify, acquire, and manage the hardware, software, and telecommunications services so that the metadata are electronically accessible;

- determine how to link or cross-reference geospatial data and metadata in order to ensure consistency, maintenance standards, and ease of query, search, and access;
- respond to questions about or orders for geospatial data.

The combination of electronic catalogue metadata with a mechanism for accessing that data is termed a data clearinghouse in the USA.

3.1 Role of a clearinghouse

In its broadest sense, a geospatial data clearinghouse attempts to connect potential data users with data producers. The clearinghouse may not provide the data directly, but it will at least inform the user of where the data might be obtained. As such, the clearinghouse may not be a central repository of information, but rather a routing station which forwards user requests to the source of the desired information. Given a set of metadata to search, the major challenge in designing the electronic clearinghouse is to devise effective data search and access mechanisms. A large number of geospatial data clearinghouses now exist around the world. Any list of them would rapidly become outdated. However, examples from various sites will be shown in the following section. The Internet addresses (URLs) of these sites have been selected for their anticipated durability and the reader is encouraged to visit them and related sites.

3.2 Search and access mechanisms

The standard free text, Boolean search mechanisms used by most WWW search engines can be used with geospatial data, but may yield strange results. The query interface to the Ordnance Survey's Glimpse index is shown in Figure 1. A full-text search of a database using the phrase 'ice and Colorado' would return records not only about ice fields located in the state of Colorado, but also Antarctic ice samples stored in freezers in Colorado. Allowing free-form text entry and Boolean searching gives the user added flexibility in finding information. By restricting searches to particular metadata fields, users who have a good understanding of the metadata structure can design their own customised search interfaces and reduce the return of spurious results. With most geospatial data catalogues, search mechanisms that deal transparently with the content, location, and temporal aspects of geospatial data are needed.

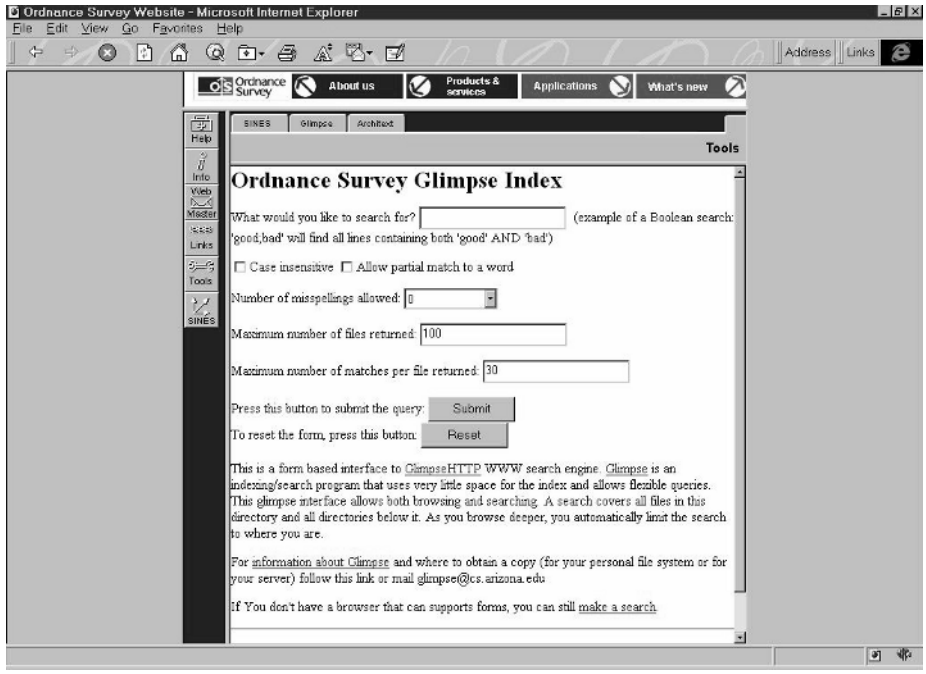


Fig 1. Glimpse interface (*http://www.ordsvy.gov.uk*).

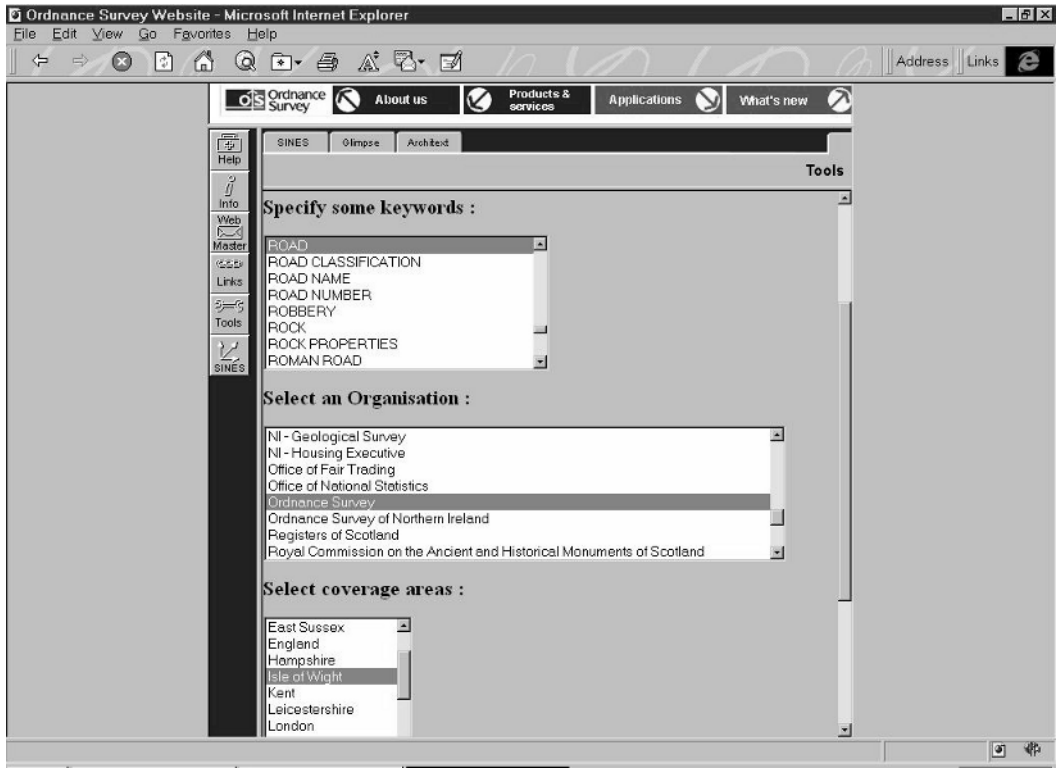


Fig 2. SINES interface (*http://www.ordsvy.gov.uk/sines-bin/SearchCGI*).

The interface to the Spatial Information Enquiry Service (SINES) clearinghouse (Figure 2) is fairly standard for geospatial data. The user selects from a list of predefined keywords (which reference the keywords in the underlying metadata), a list of data providers, and a list of pre-defined geographical areas. The use of fixed lists simplifies the search and indexing mechanisms and may help to guide the user to find the types of information that is contained in that clearinghouse. This 'pick from the list' strategy can be used to produce unambiguous results to the 'ice in Colorado' question by associating the 'ice' parameter with a set of subject keyword fields, and 'Colorado' to a location field so that the user retrieves only those records describing datasets about ice in Colorado.

The next interface adds a free-form geographical input. The user is given a choice to limit the geographical extent of the search either by a pre-defined geographical area, or by typing in the coordinates of a latitude, longitude-bounding rectangle. The interface to the FGDC clearinghouse is shown in Figure 3.

Of course very few people know the latitude and longitude extent of an area they are interested in. So a more user-friendly interface provides the user with a choice of entering coordinates or picking them off a map. This interface to the Global Land Information System (GLIS), shown in Figure 4, is tightly coupled to the underlying types of data it is serving, namely satellite imagery.

One of the most elegant interfaces is the Java-based system used by the Master Environmental Library of the US Department of Defense, shown in Figure 5(a). It includes extensive spatial search capabilities. The user can not only pick the desired area of interest from a set of reference maps but can also specify the type of spatial relationship that the dataset has with the geographical area (e.g. overlaps, entirely contained, entirely outside). The temporal nature of the catalogue data is searched by clicking on a 'temporal' tab (Figure 5(b)), while a complex series of pull-down keyword menus allow the user to specify the desired content (Figure 5(c)). The last tab shows the user the status of the data servers that are connected to the library (Figure 5(d)). The results of

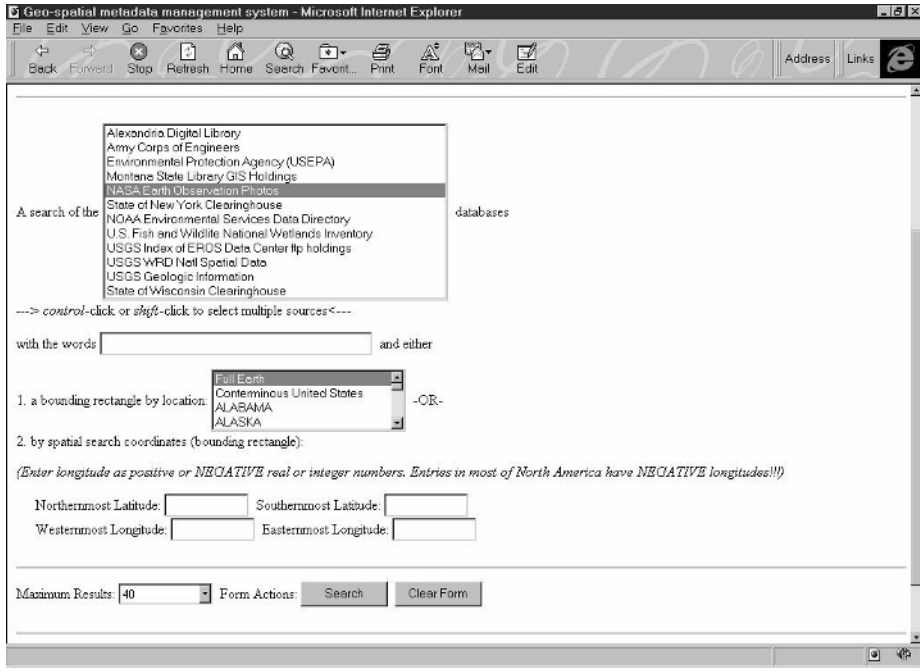


Fig 3. FGDC clearinghouse interface (*http://www.fgdc.gov/clearinghouse/index.html*).

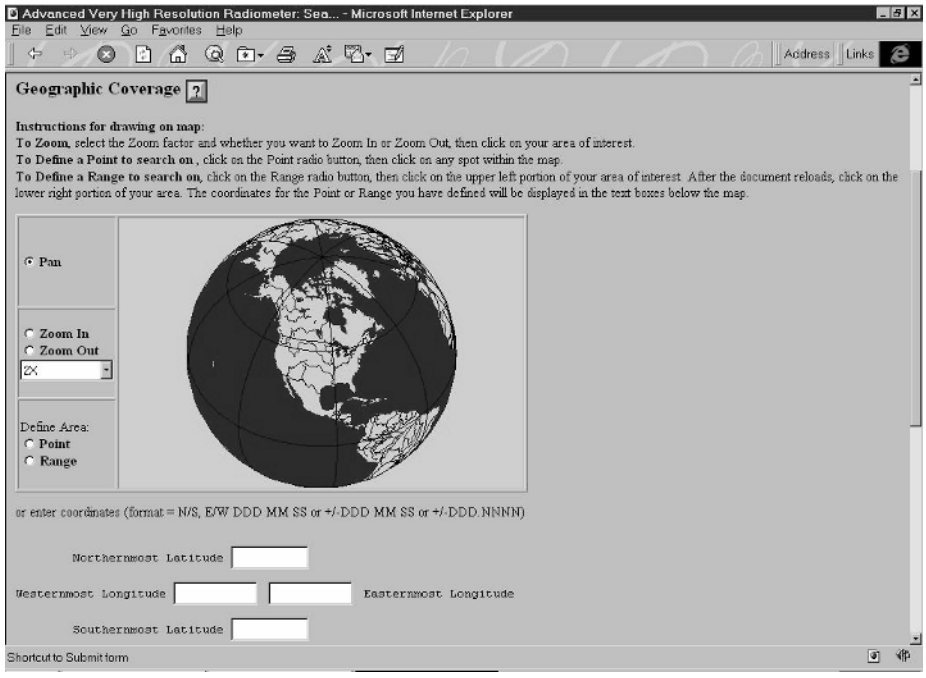
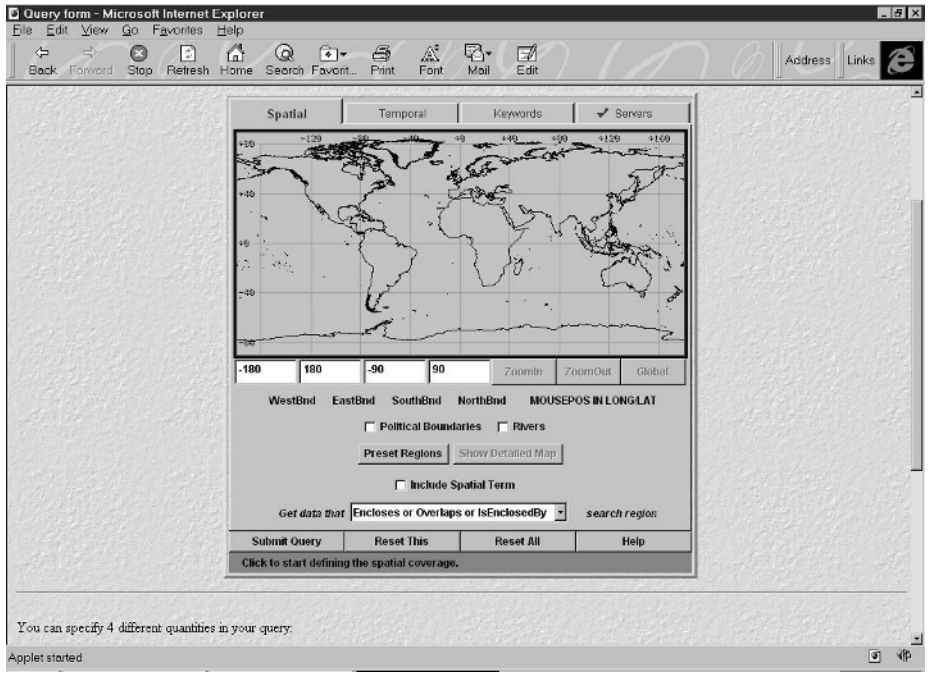
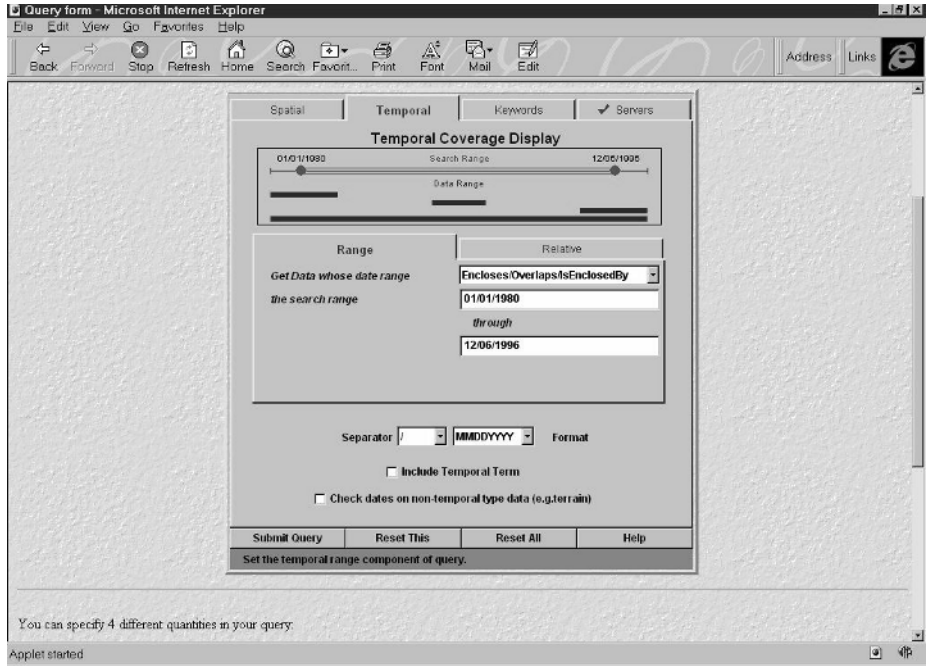


Fig 4. GLIS interface (*http://edcwww.cr.usgs.gov/webglis*).

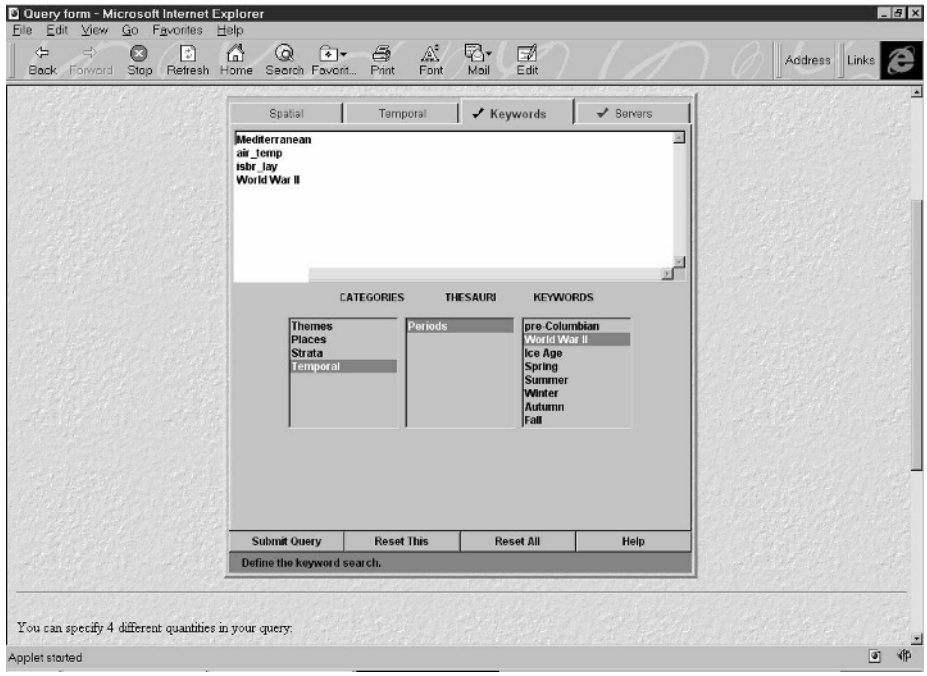


(a)

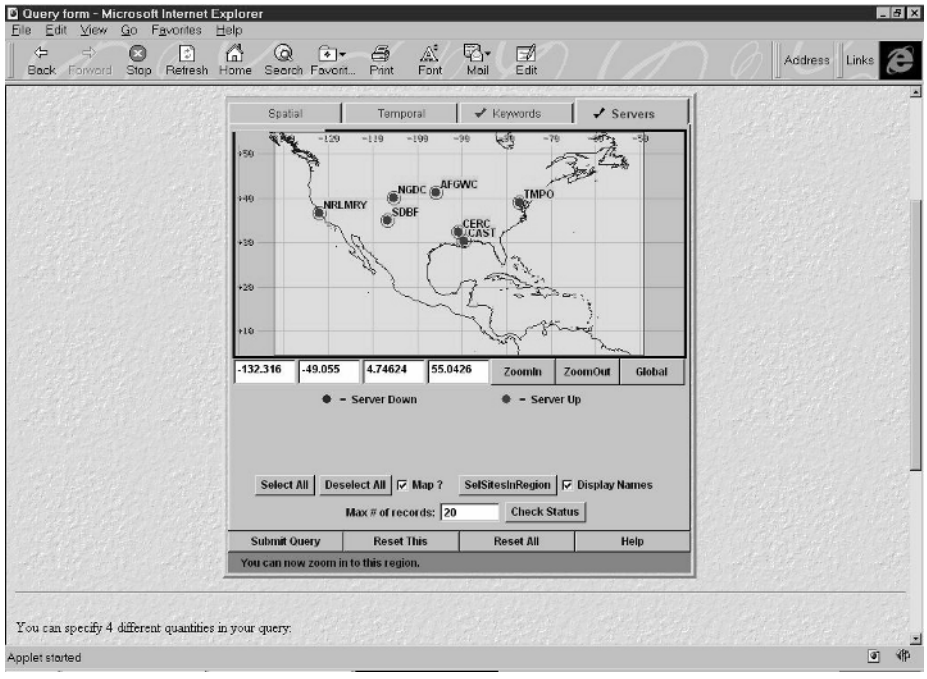


(b)

Fig 5. Master Environmental Library (MEL) interfaces: (a) spatial search capabilities (*http://www–mel.nrlmry.navy.mil*) (b) temporal catalogue data (below); (c) keyword menus (overleaf); (d) data servers; (e) query result browsing facility (page 691).

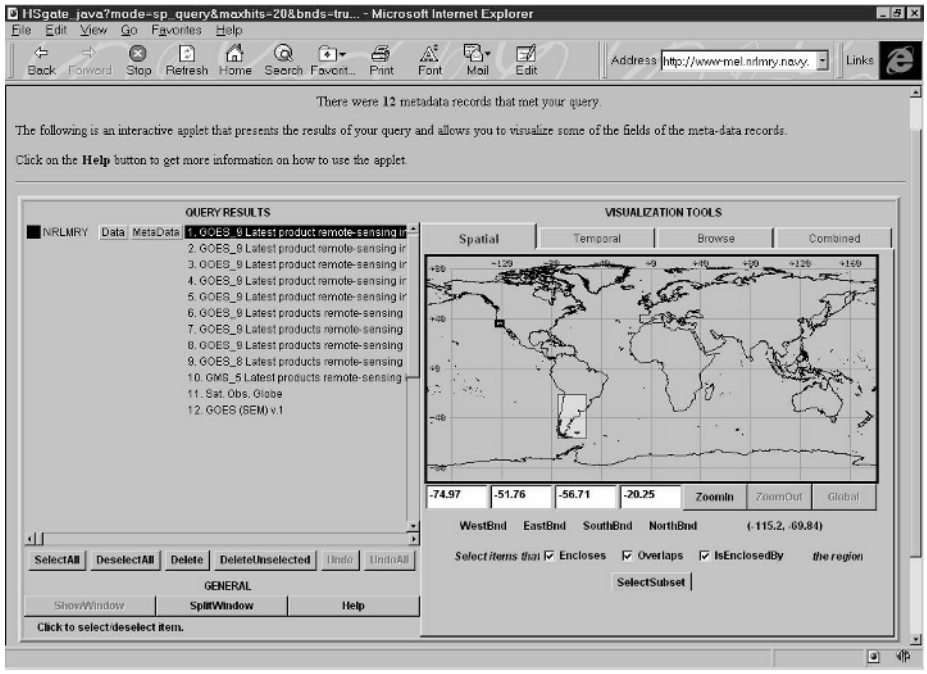


(c)



(d)

Fig 5 (continued)



(e)

Fig 5 (continued)

a query are presented in a visually-oriented interface that lets the user browse some of the available datasets (Figure 5(e)).

Advances in WWW technology and user interface design are assisting in enhancing user access to geospatial data holdings and making it easier to discover which data are available. However, determining the 'fitness for use' of a dataset is more problematic. Such determinations require both a knowledgeable data consumer and a significant amount of detailed metadata. Providing a meaningful synopsis of such metadata in a clearinghouse environment remains a challenge.

4 CONCLUSIONS

The development and use of a standardised set of metadata are the keys to organising and maintaining an organisation's investment in geospatial data and to providing information about those data holdings to data catalogues, clearinghouses, brokerages, and users. It is critical that a consensus among the numerous metadata standardisation efforts be reached on a set of core metadata elements. These core elements will most likely be stratified according to use, i.e. one set of core elements for catalogues, a second (intersecting) set for management records and a third to accompany a dataset. An agreement on common core elements will hasten the development of efficient software tools to lower the high cost of metadata entry.

The use of GIS technology continues to spread and, along with it, the demands for more and better geospatial data. The abilities to search for data and to determine their relevance are critical needs in a knowledge-based economy. Data clearinghouses provide the information to find, share, and use geospatial data more economically. The rapid advances in WWW technology are making data clearinghouses easier to use for data discovery. But we still await an effective set of tools that will allow users to determine the 'fitness for use' characteristics of any given dataset.

References

ANSI/NISO Z39.2-1994 1994 *Information interchange format*. New York, American National Standards Institute Inc.

Barton G S 1996 Multiple metadata formats from the NOAA Environmental Services Data Directory. *Proceedings, First IEEE Metadata Conference*. IEEE Computer Society Mass Storage Systems and Technology Technical Committee *http://www.computer.org/conferen/meta96/meta_home.html*

Christian E 1996 GILS. What is it? Where is it going? *D-Lib Magazine*. December 1996. *http://www.dlib.org/dlib/december96/12christian.html*

Danko D 1997 Perspectives in the development of ISO metadata standards, in *Proceedings, Earth Observation (EO)/GEO World Wide Web Workshop 97*. February 4–6, Washington DC *http://www.fgdc.gov/clearinghouse/pubs/nima/nimapaper.html*.

FGDC 1994 *Content standards for digital geospatial metadata*. Reston, Federal Geographic Data Committee Secretariat

Hakala J, Husby O, Koch T 1996 Warwick framework and Dublin core set provide a comprehensive infrastructure for network resource description. *http://www.ub2.lu.se/tk/warwick.html*

ISO 2709-1996 1996 *Format for information exchange*. New York, American National Standards Institute Inc.

ISO 15046-15 1997 *ISO Standard 15046-15, Geographic information – metadata*, Version 2.0, 20 January 1997

Johnson D, Shelley P, Taylor M, Callahan S 1991 The FINDAR Directory System: a meta-model for metadata. In Medyckyj-Scott D, Newman I, Ruggles C, Walker D (eds) *Metadata in the geosciences.* Loughborough, Group D Publications: 123–37.

Keller T, Jones D 1996 Metadata: the foundation of effective experiment management. *Proceedings, First IEEE Metadata Conference.* IEEE Computer Society Mass Storage Systems and Technology Technical Committee *http://www.computer.org/conferen/meta96/meta_home.html*

Library of Congress 1996 *The USMARC formats: background and principles*. Washington DC, Library of Congress *http://lcweb.loc.gov/marc/96princil.html*

Library of Congress 1997 *Metadata, Dublin Core and USMARC: a review of current efforts*, Discussion Paper No. 99. Washington DC, Library of Congress: 4 *gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc*

Lillywhite J 1991 Identifying available spatial metadata: the problem. In Medyckyj-Scott D, Newman I, Ruggles C, Walker D (eds) *Metadata in the Geosciences*. Loughborough, Group D Publications: 3–11

Lownsbery B, Newton J 1996 The key to enduring access: multi-organisational collaboration on the development of metadata for use in archiving nuclear weapons data. *Proceedings, First IEEE Metadata Conference*. IEEE Computer Society Mass Storage Systems and Technology Technical Committee *http://www.computer.org/conferen/meta96/meta_home.html*

NASA 1996 *Directory Interchange Format (DIF) writer's guide, version 5.0a*. Washington DC, NASA *http://gcmd.gsfc.nasa.gov/difguide/difman.html*

Newton J 1996 Application of metadata standards. *Proceedings, First IEEE Metadata Conference*. IEEE Computer Society Mass Storage Systems and Technology Technical Committee *http://www.computer.org/conferen/meta96/meta_home.html*

Renner S A, Rosenthal A S, Scarano J G 1996 Data interoperability: standardisation or mediation. *Proceedings, First IEEE Metadata Conference*. IEEE Computer Society Mass Storage Systems and Technology Technical Committee *http://www.computer.org/conferen/meta96/meta_home.html*

Weibel S, Godby J, Miller E, Daniel R 1995 *OCLC/NCSA Metadata Workshop Report*. Dublin, Online Computer Library Centre *http://www.oclc.org:5046/conferences/metadata/dublin_core_report.htm*