RSML: A PROPOSED XML BASED FORMAT AND PROCESSING LANGUAGE FOR REMOTE SENSING DATA

K.K. Mohanty
Earth Sciences and Hydrology Division, Marine and Water Resources Group,
Remote Sensing Applications Area, Space Applications Centre (ISRO),
Ahmedabad – 380 015, INDIA
e-mail: mohantykk@yahoo.com

Keywords: XML, HTML, SGML, Remote sensing metadata, spatral data products

Abstract

Remote Sensing Markup Language (RSML) – a set of XML based specifications for remote sensing metadata and operators have been proposed. The proposed specifications are based on an object-oriented decomposition of metadata requirements for remote sensing applications. An XML schema for RSML, which can be used by both data producers and application scientists, has been defined. RSML also includes specifications for remote sensing models and processing algorithm. Supporting operations among diverse data and resolution types has also been introduced. A prototype implementation that demonstrates a limited number of features of RSML has been developed. RSML is extensible and can evolve as an open standard in remote sensing.

Overview

1. Introduction
2. Current Status of Remote Sensing Metadata and Image Formats
3. Criteria for Remote Sensing Image Format
4. Remote Sensing Metadata Model
5. Criteria for Remote Sensing Processing Algorithms Standard
6. Extensible Markup Language (XML)
7. Remote Sensing Markup Language (RSML)
8. Conclusion
9. Acknowledgements
10. References

1. Introduction

Diversity in remote sensing data, models, processing algorithms and their sources has added to the complexity of earth observation application programs. The formats in which remotely sensed data and models are available or exchanged are quite diverse. The data of even a given type say passive visible and near infrared, vary widely in terms of spatial resolution, precision, projection, etc. A considerable effort goes towards interoperability of various data sets acquired from diverse sources. There is also a need to keep track of a large number of inputs, intermediate results, processing sequences and outputs, besides various metadata, e.g. gain, offset, sun angles, sensor view angle, atmospheric parameters, single as well as multi-spectral image statistics, look-up tables, etc. Currently, the choice for metadata format is quite limited. Retrieval of even commonly used parameters such as sun angle and sensor gain setting requires specialised effort.

Similarly, the image processing algorithms and earth observation models are required to operate on images of differing data and resolution types. The same image or data may also be made available under a number of coordinate / georeferencing systems. Moreover, the results need to be exchanged among users working under different system and processing environment. All these call for an open standard for remote sensing image format, operators and algorithms (Kafatos, 2002). There have been a number of efforts to address some of these issues, *e.g.* HDF (Hierarchical Data Format) from NCSA, DIMAP (Digital Image Map) from SPOT and GML (Geographic Markup Language) of OpenGIS.

This paper evaluates requirements and current scenario of metadata formats used in remote sensing. An object-oriented model of metadata requirements has been proposed from an application user's point of view. The proposed model is ideal for implementation using XML (eXtensible Markup Language) schema. The implementation is referred as RSML (Remote Sensing Markup Language), analogous to its counterparts in other scientific domains *e.g.* Chemical Markup Language (CML), Geographic Markup Language (GML), etc. (Cuthbert *et al*, 2000). The proposed schema also accommodates remote sensing processing operator, algorithm and model in the line of XSL (eXtensible Stylesheet Language).

2. Current Status of Remote Sensing Metadata and Image Formats

The various image formats currently in use in remote sensing can be broadly categorised under three groups namely i) raw data formats used by remote sensing data centers or producers *e.g.* CEOS/LGSOWG, HDF etc, ii) various graphic image formats e.g. tiff, sun raster, jpeg, gif, bmp *etc* and iii) proprietary formats of image processing software e.g. *pix* format of EASI/PACE, *lan* and *img* of ERDAS, *mips* of TNT/MIPS etc.

The raw data formats such as CEOS or LGSOWG are not self-describing, i.e. the specifications are not part of the metadata itself, rather are available as a separate document. Each of the data type e.g. passive visible and

near infrared images, SAR, altimeter etc., even when supplied from same data producer, has its own format. HDF proposed by NCSA is quite versatile, since it is self-describing and accounts for various scientific data types. HDF requires dedicated software or Application Programming Interface (API) for data parsing. It is also not specific for remote sensing metadata in an application-processing environment. SPOT 5 producers have made a very special effort to address the issues discussed here using DIMAP format (Bally, Somer, and Angleraud, 1999).

The graphic image formats fail to meet the diverse requirements of remote sensing applications and are limited in use as exchange formats among different image processing systems. Geotiff is the most successful as an exchange format, since it preserves the georeferncing information.

The proprietary image formats are specialised to address the issues of application processing environment and not the metadata specifications used by data producers. Most of the image processing (IP) software have provision for powerful user specific image processing operation e.g. Modeler from ERDAS/IMAGINE and model utility of EASI/PACE. However, there is lack of uniformity in syntax of these utilities.

3. Criteria for Remote Sensing Image Format

The design goal of the proposed remote sensing data format from an application-processing environment can be summarised as follows:

3.1 Open standard: RSML can emerge as an open specification for remote sensing, in the line similar to technologies in other application domains such as CML (Chemical Markup Language), GML (Geographic Markup Language) etc. (Murray-Rust and Rzepa, 1999; Cuthbert *et al.*, 2000). It is concerned only with metadata format and can operate with existing raw, proprietary and graphics image formats. RSML will provide a uniform syntax for producer, application scientist and end user.

3.2 Self-describing: The metadata descriptors used in the current implementation will be a part of the metadata itself and there is no need to browse through lengthy documents of data format and map them to record and fields. The entire content is in human-readable ASCII text format and can be deciphered using any XML compatible browser.

3.3 Extensible, object-oriented and hierarchical: Traditional metadata formats leave limited scope for future modification in the form of spare fields. RSML is rooted on an object-oriented model of the remote sensing metadata requirements and can be extended to accommodate newer metadata types using specialisation.

3.4 Platform independent: All data types used in RSML have a platform independent specification based on XML schema data types and hence, the entire metadata standard is platform independent.

3.5 Language neutral / internationalisation: The element names used in the specification have provision for defining alias to support different international languages. This is a feature inherently supported by XML schema.

3.6 Data integrity and validity checking: RSML defines restriction on range and type of metadata values. These restrictions can be enforced using standard and readily available validating XML parsers.

3.7 Interoperability with existing raster and vector formats: RSML is concerned with metadata format and can deal with any existing raster and vector formats for the remotely sensed data itself.

3.8 Automatic data discovery and access (including on-line access): Object-oriented, hierarchical and self-describing nature of the data format will support catalog and search services even in a distributed environment.

4. Remote Sensing Metadata Model

Data producers govern the primary format of remote sensing data and metadata. The raw data formats used by the data producers are converted to application specific formats using dedicated software. The remotely sensed data along with ancillary data and independent scene specific measurements are processed to derive value added products and information for end use. A conceptual model of the remote sensing environment is shown in *Figure 1*.

The current study is confined to a limited set of features used in remote sensing with a provision to extend the model to a larger domain. The object-oriented model presented in this study doesn't address the behavioral aspect of the system.

The metadata generation and maintenance responsibility is shared among the three major players in a remote sensing system namely, producer, and application and decision maker / end user. An object model of metadata used by data producers derived from analysis of the currently used metadata formats such as CEOS / LGSOWG etc., is shown in *Figure 2.* The metadata type of various sensors and platforms vary widely. The metadata used to describe a SAR sensor is completely different from that of a passive radiometer operating in visible and near infrared region. Similarly, the metadata model of a satellite platform is different from that of an aircraft. Such issues are addressed using specialisation / inheritance. The proposed model suggests a set of abstract sensor and producer independent objects, which are further specialised to more specific product types. The low level objects identified at the level of data producers are as follows:

Producer: Contains general information on data producer with provision for further specialisation.

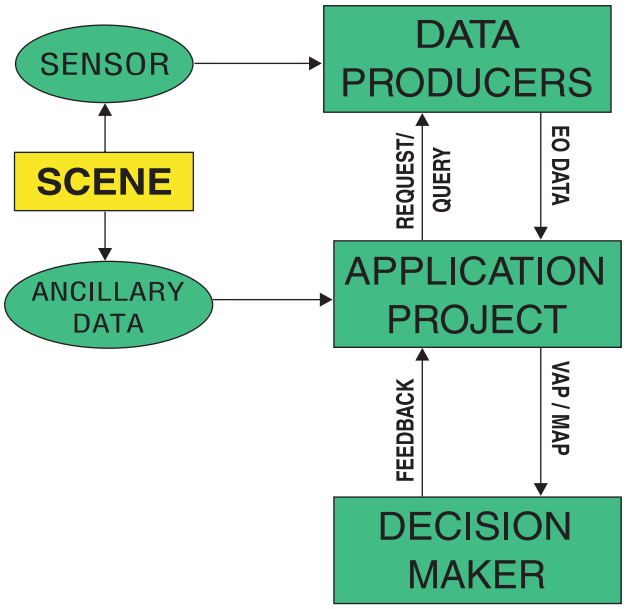

Figure 1. A conceptual model for remote sensing application system

Sensor: An abstract model of sensor containing name and reference of sensor. This needs to be specialised to a more specific Sensor object such as SAR, Altimeter *etc*.

Scene: All attributes for the target or scene *e.g.* date and time of observation, geophysical environment of imaging (sun angle, sun azimuth *etc*), georeferencing information, etc. No need for specialisation of this object is being realised at this stage.

Platform: An abstract object further specialised to more specific objects such as Satellite, Aircraft *etc*. The Satellite object accommodates all orbital details of satellite platform.

Processor: The name and version of processor, date of product generation, operator etc.

Quality: An abstract content model for data quality which requires further specialisation to various sensor specific quality measures.

Data Descriptor: An abstract object, which is further specialised to image and non-image data descriptor types. The imaging type refers to raw image storage format, dimension, data type/precision etc.

In an application-processing environment, the additional metadata relates to remote sensing products *e.g.* maps, information or value added products (including all intermediate products), processing sequence and models. The remote sensing data can be broadly categorised as raster (imaging) and vector (non-imaging). The latter includes the non-imaging type data products such as altimeter, GPS *etc*. The current analysis does not address vector data used in remote sensing. The GML object model and feature types may be adopted for this. An attempt has been made to elaborate and implement a metadata model for raster data used in remote sensing applications. Raster collection is recognised as the primary raster data type which is an aggregation of one or more raster metadata objects, multi-dimensional statistics and history. The raster constituents of a raster collection need not necessarily have the same spatial extent or data

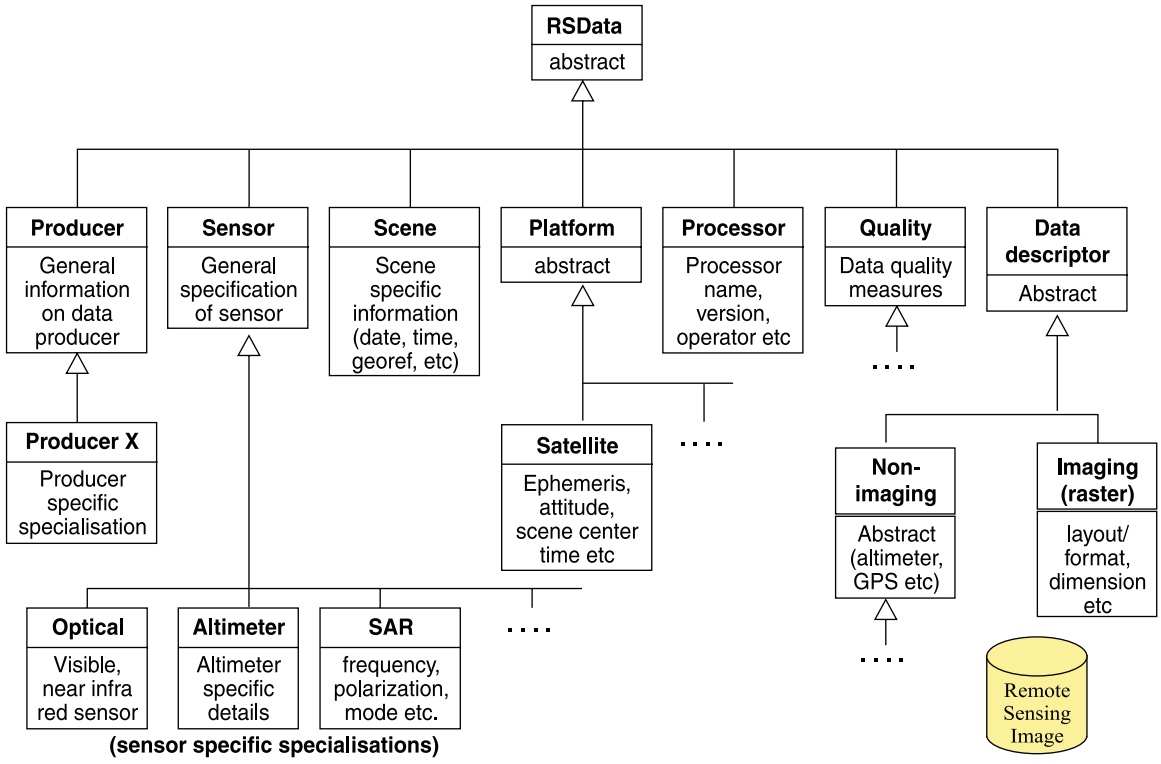

Figure 2. A model of remote sensing metadata used by data producers

type or spatial resolution. Raster object is a composition of other metadata objects such as look-up table, pseudo color table, thematic information, raster statistics, georeferencing and processing history (*Figure 3*).

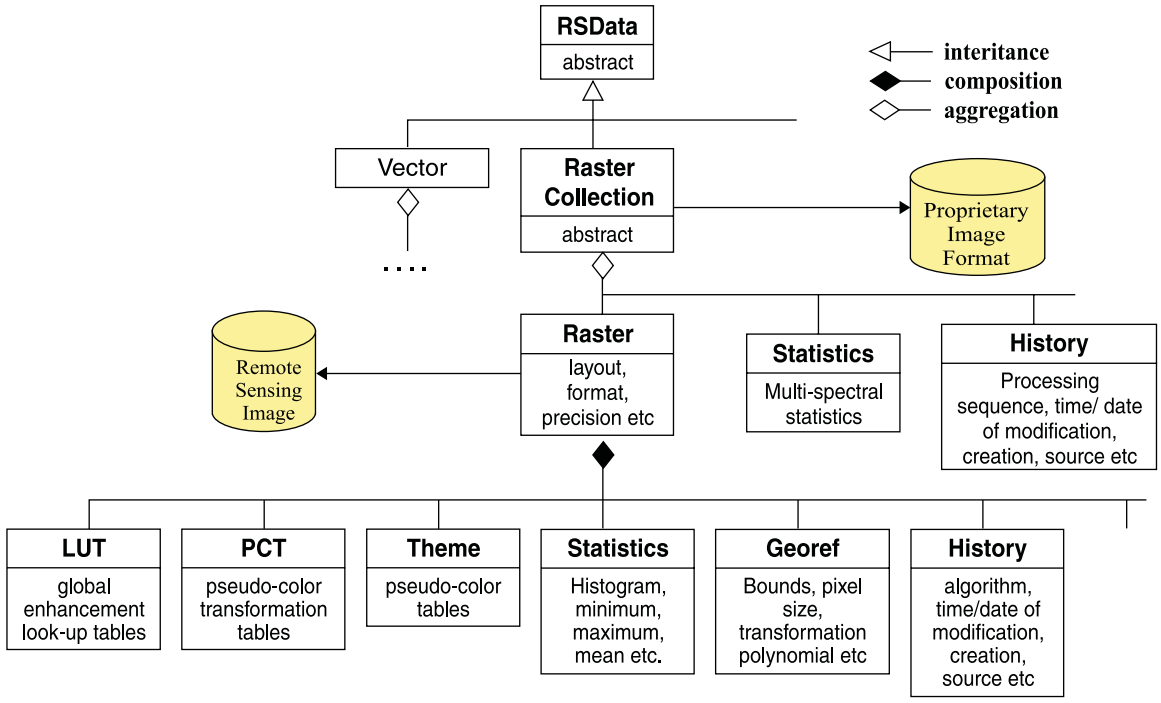

Figure 3. A metadata model for remote sensing data in application environment

5. Criteria for Remote Sensing Processing Algorithms Standard

The requirements analysis for remote sensing processing algorithms and processes primarily concerns raster (imaging) data types. The broad criteria for remote sensing processing algorithms and models are set as the followings:

- *Uniform and standard syntax:* Remote sensing operators, algorithms and models performing the same task with similar input-output data types have multiple implementations by different vendors. There is a need for uniformity in syntax of all processors. This can be achieved without disturbing the original implementations using wrappers as has been done through SQL in the field of RDBMS.
- *Inputs with different precision:* Input data types can be at different precision. This is supported even today by all implementations.
- *Inputs with different spatial resolution:* Input data types can be at different spatial resolution. Most image processing systems require inputs to be converted to identical resolution prior to joint analysis.
- *Output at any desired precision and resolution:* The output of a processing algorithm can be at any spatial resolution and precision and not necessarily as that of its input.
- *Support for distributed processing:* The inputs used by an algorithm or operator may be distributed across a network or in diverse media types.

An object-oriented model conforming to above specifications is shown in *Figure 4*. The three abstract operators, *namely* Point, Spatial and Transform, are further specialised to appropriate types. Each operator is constrained by the number of inputs and outputs. Raster algorithm refers to algorithms which use at least one raster as input.

6. Extensible Markup Language (XML)

In recent years, XML (eXtensible Markup Language), has emerged as a standard for putting structured data in a text file. XML refers to a set of rules, guidelines, and conventions, for designing text formats for storing and transmitting structured data (Bray, Paoli, and Sperberg-McQueen, 1998). XML documents are not only easy to generate and analyse, but are also unambiguous, and avoid common pitfalls, such as lack of extensibility, lack of support for internationalisation/localisation, and platform-dependency. Even though primarily meant for use in web environment, it has been used for defining many industry standards.

XML being a text tag based format, occupies more storage volume than their binary counter part. However, this is never an issue, since the metadata volume is always a fraction of actual data, particularly in remote sensing. XML is widely accepted and promoted by

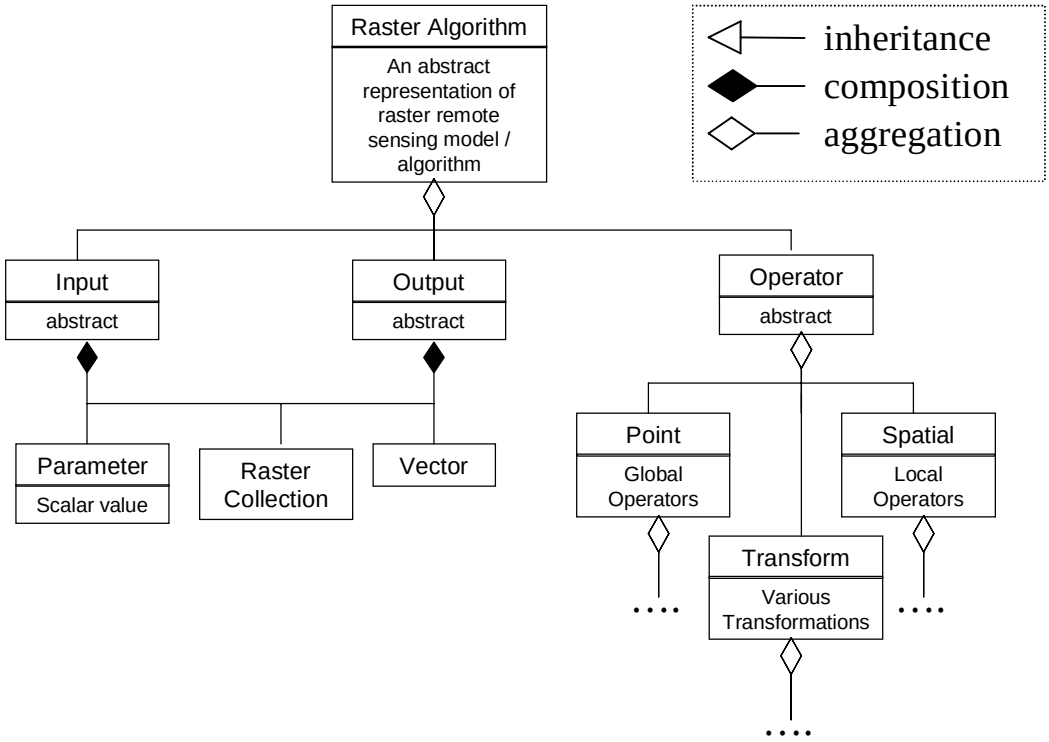

Figure 4. A model for raster remote sensing operators / algorithms

industry forum, and a number of tools and API (Application Programming Interface) are readily available for dealing with XML applications/documents.

There are two choices for defining and enforcing the structure, content and semantics of XML documents, viz. the traditional DTD (Document Type Definition) and a more recent technology - XML Schema. An XML Schema, which itself is also a valid XML document, defines the elements that can appear within an XML document, their structure, and the valid values and attributes that can be associated with any element. Schema offers many advantages over DTD, namely enhanced data types, support for user-defined data types, internationalisation, ability to derive new data types from existing ones, associating same element name with different content model *etc*. A number of XML based solutions for earth observation systems have been proposed by various workers (Smith, Passmore, Jordan, and O'Connor, 2002; Zaslavsky, 2000).

7. Remote Sensing Markup Language (RSML)

Like HTML (Hypertext Markup Language) is based on DTD defined using SGML (Structured Generalized Markup Language), RSML is defined using an XML schema based on the object models proposed earlier (*Figures – 2, 3, 4*). Two separate schemas, *namely* rsml.xsd and rspl.xsd, for remote sensing metadata format and processing algorithms are defined. Any remote sensing metadata and processing algorithm/ model conforming to RSML should refer and use these schema. RSML metadata documents can be parsed using any XML schema compatible browser such as Internet Explorer 6.0 and Netscape Navigator 6.0. Such metadata and algorithm definitions will be well-formed XML documents, whose validity can be ensured using XML schema validating parsers. Validating a schema refers to ensuring correct data types and value ranges for all metadata and predefined structure of appearance of the elements. The metadata documents can be edited or created using any standard XML editor, but in most common scenario, will be created and parsed by standard applications *e.g.* data product generator, image processing, *etc*. An application that creates and edits metadata using RSML documents has been developed using MSXML 3.0 as demonstration. The application can also parse and perform basic image arithmetic/logical operations using a RSML defined image operators.

8. Conclusion

An analysis of current scenario, issues and user requirements in metadata format in remote sensing is presented along with a XML based solution. The solution termed as RSML is presented in the form of XML schema definition. The definition is not exhaustive and more of a prototype in nature. The definition can be extended further through inheritance without discarding the base types. An analysis of requirements of a common syntax in remote sensing model and algorithm accompanied by a XML schema definition for remote sensing processing language is also presented. The proposed solution gets the benefit of XML – a technology well accepted by the

industry, and may be taken up as an open standard in remote sensing metadata format. A sample implementation of RSML is also developed.

9. Acknowledgements:

The author would like to thank Dr. S.R. Nayak, Group Director, MWRG/RESA/SAC for supporting this study.

10. References

[1] Bally, P., Somer, Y., and Angleraud, C., 1999, The Spot Image Coherence Product and Dimap: A New Format for a New Product, Proceedings Committee on Earth Observation Satellites (CEOS), 1999.

[2] Bray, T., Paoli, J., and Sperberg-McQueen, C.M., 1998, Extensible Markup Language (XML) 1.0, W3C Recommendation 10-February-1998, REC-xml-19980210, http://www.w3.org/TR/1998/REC-xml-19980210 (seen on Sept. 10, 2002)

[3] Cuthbert, A., O'Rourke, B., Keighan, E., Margoulies, S., Sharma, J., Daisey, P., and Johnson, S., 2000, Geography Markup Language (GML) v1.0, OGC Document Number: 00-029, 73p.

[4] Kafatos, M., 2002, Distributed Data Access, Analysis and Standards for Earth Science Data, A White Paper for the Strategic Evolution of ESE Data Systems (SEEDS) Public Workshop, February 5-8, 2002, Center for Earth Observing and Space Research (CEOSR), George Mason University, http://lennier.gsfc.nasa.gov/seeds/WP_Kafatos. doc (seen on Sept. 10, 2002)

[5] Murray-Rust, P. and Rzepa, H.S. 1999, Chemical Markup Language and XML Part I. Basic principles. J. Chem. Inf. Comp. Sci., 1999, 39, 928.

[6] Smith, N., Passmore, J., Jordan, C., and O'Connor, E., 2002, Appropriate Technology for Low Cost Geological mapping, GIS@Development, May 2002, Vol. VI, No. 5, pp. 40-42

[7] Zaslavsky, I., Marciano, R., Gupta, A., and Baru, C., 2000, XML-based Spatial Data Mediation Infrastructure for Global Interoperability. 4th Global Spatial Data Infrastructure Conference, Cape Town, South Africa, 13-15 March 2000.