SPATIAL DATA CONTENT STANDARDS FOR AFRICA

Antony Cooper, Bongani Majeke, Sives Govender, Kate Lance and Larry Tieszen

CSIR, CSIR, EIS-Africa, USGS/EROS & USGS/EROS
Centre for Logistics, CSIR, PO Box 395, Pretoria, 0001, South Africa
Email: acooper@csir.co.za

ABSTRACT

With funding from the United States Agency for International Development (US AID), the EROS Data Center of the United States Geological Survey (USGS/EROS) initiated a project with EIS-Africa to develop guidelines for spatial data content standards in Africa, which was executed by the CSIR. In general, data content standards include:

** Documentation specifying the information in a data set, such as: Data dictionaries, feature catalogues and classification (feature types, attribute types, attribute domain, feature relationships); Feature instances (unique, definitive versions of features); Metadata (including data quality); Data organization (eg: XML, GML); and Reference models.*

** Formal description of a model, for example using UML. Such formal descriptions are (hopefully) embedded in the data content standards.*

With the short time limits for this project and its small size, the project focused on standards for data dictionaries, feature catalogues and classification, and data organisation for them, using automated tools where possible. The project also looked at standards for feature instances. The project deliberately did not address metadata, as much work has already been done on metadata in Africa.

We identified and assessed 170 candidate standards from various sources (eg: ISO/TC 211, Open GIS Consortium, FGDC, Zimbabwe and Standards South Africa). From these we selected several to compare their feature types (classes). The project included developing guidelines for data content standards in Africa. This presentation will describe the work done during the project and present our results.

1. INTRODUCTION TO THE PROJECT

Early in 2004, the EROS Data Center of the United States Geological Survey (USGS/EROS) initiated a project with EIS-Africa to develop guidelines for data content standards for geospatial data in Africa. The project was funded by the United States Agency for International Development (US AID). The project was awarded to the CSIR (which contributed additional funding for the project), and was executed jointly by three of its units: the Satellite Applications Centre (SAC), the Centre for Logistics, and the Unit of Water, Environment and Forestry Technology (Environmentek).

Information on the project was circulated on the SDI Africa Discussion List of the Global Spatial Data Infrastructure Association (GSDI), on the EIS Africa mailing list, and to the delegates that attended the 5th African Association of Remote Sensing of the Environment (AARSE) Conference, held in Nairobi, Kenya, in October 2004. Presentations during the project were made at several meetings in South Africa and Kenya, and the results were presented at the meeting of the Committee on Development Information (CODI) of the the United Nations Economic Commission for Africa (UN ECA) in Addis Ababa, Ethiopia, in April 2005. These results will provide useful inputs for the Working Group on Standards, established by CODI's Geoinformation Sub Committee (CODI-Geo) in August 2004.

2. INTRODUCTION TO DATA CONTENT STANDARDS

An expansion of formal definitions for *data content standard*, *application schema*, *conceptual schema*, and *universe of discourse* [GOS 2004], provides the following definition of a data content standard:

> *A standard that specifies what information is contained within a geospatial data set, and provides a formal description of a model that defines the concepts of a view of the real or hypothetical world that includes everything of interest, for data required by one or more applications.*

From this definition, data content standards include:

- **Documentation specifying the information in a data set**. This includes:
    - **Data dictionaries, feature catalogues and classification**. These define the types of geographical features (ie: the *classes* or *feature types*) one would find in a data set, together with their *attributes* (types and domains) and other peculiarities, enabling a shared understanding of the contents of the data set. They can also include conceptual relationships between feature types (eg: that an instance of the feature type *'bridge'* can carry an instance of the feature type *'road'* over an instance of the feature type *'river'*). It is critical that they include proper definitions to differentiate between feature types, and not depend merely on the label (name) attached to the feature type.
    - **Feature instances**. The instance of a feature in a data set represents a discrete phenomenon in the real (or imaginary) world – that is, something specific out there that is modelled in the data set. The instance normally has coordinates and may be portrayed on a map by a particular graphic symbol. Standards for feature instances specify the unique, definitive versions of features, eg: the draft South African standard, SANS 1876 *Feature instance identification standard*.
    - **Metadata**. *"Data about data"*, metadata describes the structure and content of a data set, and how to access and use the data. Metadata encompasses issues such as data quality (eg: accuracy, completeness, lconsistency, currency and lineage), data schemas, spatial referencing systems, constraints on the use of data, and contact details for those responsible for the data set. Metadata enables data discovery, determining fitness for use, data access and data transfer.
    - **Spatial representation**. This is how locations are represented, either through coordinates (given within the context of a spatial referencing system) or geographical identifiers (eg: a town's name).
    - **Data organization**. This encompasses the logical description of the data set, using formal languages such as the eXtensible Markup Language (XML) or its implementation for geographical information, ISO 19136, *Geography Markup Language (GML)*.
    - **Reference models**. These provide conceptual descriptions of data sets, using standards such as ISO/IEC 10746, *Open Distributed Processing – Reference model (RM-ODP)*, with its five Viewpoints (Enterprise, Information, Computation, Engineering and Technological).
- **Formal description of a model**, for example using the Unified Modelling Language (UML). Such formal descriptions are (hopefully) embedded in the data content standards and hence for the purposes of this project, we needed to only record their presence or absence and use them to understand the content of the standards, rather than analyse the models *per se*.

There are already many initiatives addressing metadata for geospatial data in Africa, and with the short duration of this project and its small size, its sponsors felt that the focus should be on other data content standards for geospatial data. This project focused primarily on standards for data dictionaries, feature catalogues and classification, and the data organisation for them.

A *data dictionary* is an unstructured collection of feature types, while a *feature catalogue* is a structured collection of feature types (eg: as a hierarchical classification), and hence easier to use. For our purposes, the terms *feature catalogue* and *classification* are synonyms. Typically, a feature catalogue would be constructed from a data dictionary, perhaps as a *profile* (subset) of the data dictionary.

Desireable characteristics of such standards include the flexibility to customise the structure of one's implementation, without compromising the ability to exchange the data with others using different structures. Further desired charactistics include: complying with policies and other standards; understanding the limitations of the data; fulfilling any reporting mechanisms; and being understandable and accessible.

3. POSSIBLE DATA CONTENT STANDARDS

3.1 Standards generating bodies

There are many standards generating bodies around the world, some of which have an international scope while others have a more localised scope; some are open organisations developing standards through consensus, while others are subscription-based industry standards groups. Of particular relevance are:

- *ISO/TC 211*, the Technical Committee (TC) of the International Organization for Standardization (ISO)developing standards for *Geographic Information/Geomatics*, the ISO 19100 series.

- The *Open Geospatial Consortium, Inc (OGC)*, an industry standards group working closely with ISO/TC 211.

- The American *Federal Geographic Data Committee (FGDC)*, which developed standards that have been widely used around the world, some being pre-cursors to ISO 19100 standards.

Within Africa, we sourced candidate standards from national standards generating bodies in Zimbabwe and South Africa. Currently, the Southern African Development Community Cooperation in Standardization (SADCSTAN) is adopting standards for geographical information from South Africa and ISO/TC 211.

3.2 The data content standards considered

USGS/EROS provided the project team with a collection of candidate data content standards from various standards generating bodies. We also considered standards from organisations in which the project's team members are active, and looked for sources for such standards on the Web. We attempted to find candidate standards in French, but the only ones we could identify are those standards developed by ISO/TC 211 that Canada has translated into French, such as ISO 19115, *Information géographique – Métadonnées*.

We compiled a spreadsheet of the 170 candidate standards we identified (see Cooper & Majeke [2005] for the full list), recording for each details we could extract, such as the standard's name, its source, its official identifier, its date published, and its status. We also tried to determine whether or not it was a data content standard and whether or not it was relevant for the project. Some standards are available only at a price and to access others one needs to subscribe to the standards, which in some cases is a long and complex process – we limited ourselves to those that are widely available. From these, we made a selection for detailed comparison with a reference standard.

4. COMPARISON OF FEATURE TYPES

4.1 A reference standard

To facilitate assessing the contents of standards for data dictionaries and feature catalogues, we selected SANS 1880, the *South African Geospatial Data Dictionary (SAGDaD)*, as a reference standard, largely because of our personal experience with it. SAGDaD is the South African implementation of ISO 19110, *Geographic information – Methodology for feature cataloguing*. SAGDaD focuses on the core data types likely to be transferred between users, and currently defines 81 feature types and 29 attributes, with the attributes being defined uniquely across all the feature types. One of these attributes is Enumerated Type, which refines a feature instance's feature type to provide subtypes or subclasses.

4.2 The standards compared

The process of comparing the feature types from these standards was more difficult than we had initially anticipated. Different countries use terminology differently, and have different legal frameworks and other factors that influence the labels they apply to feature types and how they structure feature catalogues. Not all the standards had definitions readily available, and it was not feasible to trawl through in fine detail all those definitions that were available, to determine the subtle nuances used to create each feature type.

The classification of digital geographical information is a subjective process because people observe different properties in features and require information about the features to different levels of detail [Scheepers *et a*l 1986]. There is a grey area between feature types and attributes: what constitutes a feature type in one data dictionary or feature catalogue, could constitute an attribute of a feature type in another.

Finally, some of the data dictionaries and feature catalogues cover the whole gamut of digital geographical information (though perhaps focusing only on the core feature types), while others focus on a narrow aspect of digital geographical information (though providing feature types to a greater level of detail). Nevertheless, we believe that the resultant comparison is useful, highlighting core feature types.

The feature types of SAGDaD and of the chosen standards were compared in a spreadsheet, whereby similar feature types from different standards were put in the same row. We selected three FGDC and ten Zimbabwean standards for comparison with SANS 1880 [Cooper & Majeke 2005].

Some of the feature types from the candidate standards matched up well with the corresponding feature types in SANS 1880, but there were those that did not. SANS 1880 provides a general purpose data dictionary principally for core data sets, while some of the others provide detailed feature types for their particular application domain. SANS 1880 also contains refinements of its feature types through its Enumerated Type attribute, but these attribute values were not included in the integrated feature catalogue. It is probable that many of those unmatched feature types in the other standards would match one of these feature type and attribute combinations in SANS 1880.

5. ADVICE ON DATA CONTENT STANDARDS

5.1 Collection criteria for feature types

ISO 19110, *Geographic information – Methodology for feature cataloguing*, does not contain a feature catalogue *per se*; it specifies the structure of a compliant feature catalogue. Unfortunately, it does not address the collection criteria for compiling a feature catalogue – that is, any requirements or guidelines for how to identify and define a "good" feature type or feature catalogue. ISO 19126, *Geographic information – Feature data dictionaries, feature catalogues and registers*, will most likely use a profile of ISO 19110's conceptual model for feature catalogues, define the conceptual model for a data dictionary, and specify how these form the basis for establishing and managing registers for data dictionaries and feature catalogues. Nevertheless, it will also not address the collection criteria.

Collection criteria can be very complex and dependent on the field of application, as well as on one's cultural, linguistic, legal and political framework. Hence, it is likely that one will have to deal with a variety of data dictionaries and feature catalogues, with translation tables between them. Unfortunately, it can be tedious to set up such translation tables. Normally, data will be transferred from a user that has deeper (or the same) knowledge of the data being transferred than the recipient has – an individual feature is more likely to be transferred from the user that would classify the feature more precisely [Cooper 1993]. Hence, such mappings are invariably one-to-one or many-to-one, and can be done automatically.

Unfortunately, it is far too common to categorise things for the wrong reasons – that is, use invalid collection criteria. Typical mistakes identified by Cooper [2003] include:

- **Using a quantitative measure to differentiate feature types**: a small change in the measure that then crosses a threshold will necessitate reclassifying a feature, such as when towns are classified on the basis of their population. Unless there are obvious breaks in the range of numeric values, these feature types could also appear to have artificial boundaries.

- **Overloading a feature type**: this is very common, and occurs when one feature type is used to convey several different meanings (often independent), which could cause confusion or which could render the feature type invalid when one has better knowledge. The identification of a feature type should be based on only the most important set of its characteristics.

- **Assuming there is only one valid categorisation**: rather than overloading a feature type, one should consider whether or not there are two or more perspectives of the same data, and develop a taxonomy for each. For example, a post office could be viewed as being of the feature types *'building'* (for the physical building), *'post office'* (as part of the postal network), and *'government'* (for land use zoning of the erf).

- **Categorising the symptoms, not the causes**: it is easiest to start by categorising the effects (symptoms, or superficial aspects) of one's subject of interest, while it is much more useful to categorise based on the causes (fundamentals).

- **Making the categorisation dependent on its encoding**: this occurs typically with an hierarchical classification, where the number of feature types on a level is limited by how they are encoded (eg: by having a structured code with a single digit for the feature types at each level). There are many classifications that fall into this trap. This places an artificial limit on the number of feature types, or results in superfluous feature types being created. This often manifests itself in the 'need' to have a round number of feature types.

- **Assuming there is a perfect categorisation**: it is very easy to get into *"analysis paralysis"* trying to develop the perfect standard – and then once it is released into the real world, discover its shortcomings. Hence, one should expect to revise whatever categorisation is developed, especially based on feedback

from users. It is best to test the proposed categorisation as quickly as possible in the real world.

5.2 Developing data content standards

African users of digital geographical information should not be mere passive receptors of technologies (such as standards) from other parts of the world, but should be active contributors to their development. They need to play active roles in planning and developing standards, to ensure that the relevant standards are appropriate for African conditions (eg: being viable and affordable), and that they meet the needs of Africa.

A study initiated by the German Institute for Standardization (DIN), found that companies actively involved in standards development incur lower costs when the standards are implemented [DIN 2000]. They get early access to current technologies and thinking (insider knowledge), are able to assert their interests in the standardization process (getting desired contents included and undesired contents excluded), and lower the economic risk and costs of their research and development. Some key findings include:

- The savings from internal (company) standards are even greater than those from industry-wide standards;
- The positive effect of standards on communication between departments within the organisations was rated significantly higher than the effect on production costs;
- The longer the lifespan of the products, the greater the relevance of the standards, particularly industry-wide standards; and
- Using standards reduces liability risk.

Involvement in international standardisation efforts provides access to current technologies and thinking, and the possibilities to take technical or managerial leadership roles in national and international standards development. However, it can be expensive to attend the meetings of international standards generating bodies, particularly as they rarely meet in Africa, and funding is not always readily available. Indeed, some delegates participating in ISO/TC 211 meetings to pay some of the costs out of their own pockets.

Local standards need a massive investment to support their implementation, because of the small local market available to support the standard. Local vendors have limited resources, and don't necessarily have access to the source code for the products they sell, making it very difficult for them to get their products adapted to support local standards.

However, it is feasible to develop local profiles of international data content standards, as the products that support the international standard will automatically support the profile. Local profiles make the international standards easier to use, as they cater for local conditions. It should also be feasible to develop local implementations of an international data content standard, and the products that support the international standard should have the 'hooks' to accommodate the local implementation, which should be available in a flat file or a simple data base format. Again, a local implementation caters for local conditions.

In addition, African experts need to get involved in the activities of standards generating bodies (such as ISO/TC 211), to influence the development of international standards to ensure they meet Africa's needs – that way, the commercial GISs will support Africa's needs off the shelf, without needing to develop local standards. Participation in ISO/TC 211 can be done through national bodies (six African countries are members of ISO/TC 211), or through the 24 Class A Liaisons to ISO/TC 211, such as the International Association of Geodesy (IAG), the International Cartographic Association (ICA), the International Federation of Surveyors (FIG), the International Society for Photogrammetry and Remote Sensing (ISPRS), or the International Steering Committee for Global Mapping (ISCGM). Most importantly, the UN ECA is now a liaison to ISO/TC 211.

Participation in the activities of ISO/TC 211 can be done successfully via email – a very good example is that of the Czech Republic, which has made useful contributions to many standards without being represented at ISO/TC 211 meetings. However, such participation still requires much effort, as there is a high barrier of entry in the complexity of standards and one needs to dedicate the time to read the drafts of standards and supporting documentation, so that one can comment usefully on them – one needs to take care over the comments one provides on a standard to ensure that they are comprehensible and valid.

5.3 The perfect standard

There are some who believe that a 'perfect' standard can be developed by a project team purely from a theoretical basis and using their skills and experience. However, no matter how much diligence is applied to writing a new standard, invariably many problems will be discovered as soon as anyone attempts to use the standard. Hence, it is critical that the standard be tested as early as possible in the development cycle, to validate the approach being taken and to highlight key issues that the developers might have overlooked.

It is also all too easy to assume that once a standard has been approved and implemented and the relevant training given, it will be used properly by those who should use it. All implementations need to be followed by assessments of their implementation, to identify and understand:

- Non-compliance – and hence, possible remedial actions to ensure compliance, such as revising the standard;
- Implementation problems, such as software incompatibilities, missing data, onerous requirements, conflicts with other standards, or errors or ambiguities in the standard;
- Costs of implementing the standard – and hence, whether or not the standard is economically viable;
- Additional training and/or training materials needed; and
- Further standards that need to be developed.

6. RECOMMENDATIONS ON WHICH STANDARDS TO USE

As a standard for feature catalogues, ISO 19110 has some limitations, particularly relating to providing mechanisms to accommodate explicitly cultural and linguistic adaptability (CLA), that is, *"the ability for a product, while keeping its portability and interoperability properties, to:*

- *be internationalized, that is, be adapted to the special characteristics of natural languages and the commonly accepted rules for their use, or of cultures in a given geographic region; and*
- *take into account fully the needs of any category of user"* [ISO/IEC TR 11017].

Such adaptability is particularly useful in multi-lingual environments, which would apply to most, if not all, countries in Africa.

Nevertheless, we recommend that any feature catalogues used should conform to ISO 19110. Such a feature catalogue can still have feature types with labels (names) in multiple languages, implemented as aliases. Currently, the limitations to ISO 19110 are unlikely to affect most users of feature catalogues – few feature catalogues use the optional constructs provided by ISO 19110, such as feature operations and feature associations. In addition, ISO 19110 will continue to be maintained and enhanced by ISO/TC 211.

Unfortunately, it is not possible to recommend one, definitive data dictionary or feature catalogue to be used for all digital geographical information by all users across Africa under all circumstances. It is probably more effective to use a widely used feature catalogue that meets most of one's needs (adding more detailed feature types for in-house use, if necessary), than to try seek out that arcane feature catalogue that matches one's needs perfectly. The data dictionary or feature catalogue one should use depends on a number of factors:

(1) Is one using one's GIS in a narrow field, where one requires detailed feature types? Under such circumstances, it is best to use a specialist feature catalogue for that domain.

(2) Is one using one's GIS for general work, using mainly core data sets? Under such circumstances, it is best to use any general purpose data dictionary (eg: SANS 1880) likely to be used by one's data providers, though structured as a feature catalogue.

(3) Does one have one dominant client for one's data? Under such circumstances, one should use one's client's data dictionary or feature catalogue.

(4) Is no suitable data dictionary or feature catalogue readily available? Unfortunately, developing a data dictionary or feature catalogue can be very demanding (of expertise and time), while the user needs one immediately. Invariably, this results in a "quick fix" feature catalogue being developed, with subsequent long term problems with its use. Under such circumstances, one should network with one's peers to find a suitable data dictionary or feature catalogue.

(5) Are there too many data dictionaries and feature catalogues available? This is probably fairly common, with the added problems of there not being one data dictionary or feature catalogue that really meets one's needs, and there being mismatches between the various candidates. Under such circumstances, one should network with one's peers to get these data dictionaries and feature catalogues harmonized.

Unfortunately, developing or harmonising data dictionaries and feature catalogues and can a time consuming and tedious process.

7. CONCLUSIONS

We have provided an introduction to data content standards for digital geographical information in this presentation, and summarised our assessment of the 170 standards that we identified as candidates for further examination. Of these, we selected 14 standards containing data dictionaries or feature catalogues, and compared their feature types. We have also provided some advice and recommendations on data content standards (particularly for data dictionaries and feature catalogues) for Africa.

We trust that this presentation provides some useful guidelines for data content standards for Africa, and that it can be used as a basis for bringing some degree of harmonization to the data dictionaries and feature catalogues used in Africa, and to disseminating the better data dictionaries and feature catalogues to new users of digital geographical information. We also trust that this project reported on here, will serve as a basis for future research on data content standards, not only in Africa.

8. ACKNOWLEDGEMENTS

We would like to acknowledge the support of US AID, USGS/EROS, EIS-Africa and the CSIR, that made this project possible. We would also like to acknowledge the initiative of Kate Lance of USGS/EROS who initiated the project and provided us with much insight and input, as well as the contributions of Doug Nebert (FGDC), Andrew Terhorst (CSIR Satellite Applications Centre), Peter Schmitz (CSIR Centre for Logistics) and all the others who have provided us with inputs. Finally, we would like to acknowledge the support of the Global Spatial Data Infrastructure Association (GSDI) for allowing us to use their SDI Africa Discussion List.

9. REFERENCES

[1] [Cooper 1993] Cooper AK, June 1993, *Standards for exchanging digital geo-referenced information*, unpublished MSc thesis, University of Pretoria, South Africa, 247 pp.

[2] [Cooper 2003] Cooper AK, 2 October 2003, *Thoughts on categorising bloodstain patterns*, unpublished CSIR report (CSIR icomtek document number: 0442-0001-701-A1) submitted to the European Network of Forensic Science Institutes (ENFSI), as part of the consideration to establish an Expert Working Group for Bloodstain Pattern Analysis (BPA), 3pp.

[3] [Cooper & Majeke 2005] Cooper AK and Majeke B, 31 March 2005, *Guidelines for data content standards for Africa*, contract report for Eros Data Center (EDC) of the United States Geological Survey (USGS) and EIS-Africa, document number CR-2005/05 SAC, 14pp + 2 Annexes.

[4] [DIN 2000] DIN, April 2000, *Economic benefits of standardization: Summary of results*, Beuth Verlag, 39pp.

[5] [GOS 2004] Geospatial One Stop, July 2004, I*nformation Technology – Geographic Information Framework Data Content Standards: Part 0 – Base Standard [draft]*, Information Technology Industry Council.

[6] [ISO 19110] ISO 19110:2005, *Geographic information – Methodology for feature cataloguing*.

[7] [ISO 19115] ISO 19115:2003, *Geographic information – Metadata*.

[8] [ISO 19126] ISO 19126, January 2004, *Geographic information – Feature data dictionaries, feature catalogues and registers [Committee Draft]*.

[9] [ISO 19136] ISO 19136, February 2004, *Geography Markup Language [Committee Draft]*.

[10] [ISO/IEC 10746] ISO/IEC 10746, *Open Distributed Processing – Reference model*.

[11] [ISO/IEC 11017] ISO/IEC TR 11017:1998, *Information technology -- Framework for internationalization*.

[12] [SANS 1876] SANS 1876, 2005, *Feature instance identification standard [Draft South African Standard]*, Standards South Africa.

[13] [SANS 1880], SANS 1880, May 2003, *South African Geospatial Data Dictionary (SAGDaD) and Its Application [Committee Draft]*, Standards South Africa.

[14] [Scheepers et al 1986] Scheepers CF, Van Biljon WR & Cooper AK, September 1986, *Guidelines to set up a classification for geographical information*, NRIMS CSIR Internal Report I723, 12 pp.

[15] [UML 2003] *Unified Modeling Language™ (UML®)*, Version 1.5, March 2003, Object Management Group™ (OMG™). Available on 28 March 2005 from: http://www.omg.org/technology/documents/formal/uml.htm

[16] [XML 2004] *Extensible Markup Language (XML) 1.0 (Third Edition)*, 4 February 2004, World Wide Web Consortium (W3C). Available on 28 Marc h 2005 from: http://www.w3.org/TR/2004/REC-xml-20040204/