United Nations E/CONF.101/56

Economic and Social Council

Distr.: General

21 May 2012

Original: English

Tenth United Nations Conference on the Standardization of Geographical Names
New York, 31 July – 9 August 2012
Item 12 of the provisional agenda*
Terminology in the standardisation of Geographical Names

Feature Types for Global Gazetteers

Submitted by Australia**

* E/CONF.101/1.

** Prepared by: Laura Kostanski, Rob Atkinson and Paul Box (Australia), Commonwealth Scientific and Industrial Research Organisation (CSIRO) Bill Watt (Australia), Chair Committee for Geographical Names of Australasia

Summary

Gazetteers may list names of geographic locations representing a wide range of physical or administrative feature types. To understand these names, it is necessary to specify what type of feature is being referred to. This leads to a range of issues around support for multiple languages, specific domain of use for a gazetteer and how finely feature types are differentiated. When integrating data from multiple sources different feature type classifications increase the complexity of any interpretation or searching process. Most data about human and natural processes is geographically referenced, usually by some form of identifier such as a town, administrative area, address etc. Interpreting these references is a significant and pressing need for better using information to support decisions, and gazetteers are the source of the information needed to interpret these references.

In partnership with the Office of the UN Chief Information and technology Officer (CITO) and AusAID, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) has been working with the Committee for Geographic Names of Australasia (CGNA) and the Badan Informasi Geospasial (BIG) - the Indonesian National Geospatial information Agency, to pilot a gazetteer framework that can be used as global best practice as well as globally accessible persistent infrastructure. The framework, which allows users to discover the meanings of toponyms or identifiers found in geographically referenced datasets, is being developed through a project entitled the UNSDI Gazetteer Framework for Social Protection in Indonesia. The project plan and research overview are provided in another paper submitted for the Conference and further details can be located at www.csiro.au/gazetteer.

In the course of developing a prototype gazetteer framework for implementation in Indonesia, the research team has encountered a number of issues which should prove to be of interest for UNGEGN members. This paper deals with the conundrum of feature types- in particular the issues associated with mapping multiple and varied feature types in a linked gazetteer structure. Examples of existing feature catalogues are provided, along with descriptions of the benefits and drawbacks of current feature type classification methods. The research team proposes to UNGEGN members the development of a Special Committee on Feature Types (perhaps as a sub-committee of the Working Group on Toponymic Data Files and Gazetteers, or the Working Group on Toponymic Terminology), which will bring benefit by feeding input from toponymic experts into the larger UN Spatial Data Infrastructure initiative (UNSDI) being driven by the UN Geographic Information Working Group (UNGIWG). The proposed benefits of the development of a feature type and classifications list are extensive, and include the ability for naming authorities to identify areas for increasing the scope of their gazetteer data collection methods (as discussed in the *Four Faces of Toponymic Gazetteers* paper submitted by Australia for this conference).

The problem

For the purposes of explaining the applications of place names, and to facilitate data access and analysis, a gazetteer classifies features into distinct types- for example *mountains* and *waterways*. Up to now feature type categories for official gazetteers have been developed independently by jurisdictional naming authorities. Other datasets that act as gazetteers, e.g. postcodes or census districts, also have an implicit feature type, typically applying to the whole data set. Thus, there exist as many feature type classification schemes as there are gazetteers.

There is a sound rationale behind having varied feature types for each gazetteer- for example place naming is a distinctly cultural process and the definitions of features in the landscape will necessarily vary between jurisdictional naming authorities. However, we live in an interconnected and rapidly changing world with increased availability of data, network connectivity and GIS systems. We progressively need and expect effective cross-jurisdictional cooperation to address large scale pressures of social and environment protection. The creation of a gazetteer framework to support this interlinking has brought to the fore the long-standing issue of how to match feature types between disparate systems.

Integration of multiple gazetteer sources immediately raises several issues however:

- If a source gazetteer is classified using a coarser set of terms than the common list, any mapping is going to be ambiguous;
- Cultural and political sensitivities may result in difficulties in using a finely nuanced common term;
- The amount of effort required to reconcile and map many detailed source feature type lists to a common list may be large;
- As the common list evolves, for example a term is split into several related terms, the many mappings in use would need to be revisited;
- It will be difficult to develop and maintain a comprehensive multi-lingual aspect of a large common list;and
- Users would be overwhelmed by a comprehensive, finely nuanced and potentially overlapping set of reported feature types.

For the reasons listed above, we explore two approaches to resolving the feature types issue: the first is a proposal to develop a comprehensive feature type catalogue, the second is to develop a set of feature type classifications to which feature types from source gazetteers can be referenced and linked.

A common-list approach

The task of selecting or developing a perfect common list with no ambiguities would be daunting. Several broad-scope gazetteers have chosen or developed such lists, raising significant issues with evaluation and adoption (e.g. Alexandria Digital Library, Geonames.com, Yahoo gazetter, Esri gazetteer). One of the most prominent feature type lists is that developed by the Alexandria Digital Library (ADL available from http://collections.alexandria.ucsb.edu/adl_gazetteer/metadata.html). The ADL is an online gazetteer which compiles data from the United States Geological Survey gazetteer, combined with other official and unofficial gazetteers from around the world. At present there are approximately 5.3 million records in the ADL which use one of 1046 lead-in feature type descriptions.

During the development of the ADL, there was a recognised need to create definitions for feature types, to encapsulate the range and breadth of variation between gazetteer classification systems. In 2000 the ADL team sampled feature types from each of the available online gazetteers (approximately 7) and discovered that they could create a Thesaurus of feature types which had a 5:1 ratio of lead-in vocabulary to preferred terms (209 preferred terms and 978 lead-in terms). Within the thesaurus *lead-in vocabulary* is defined as the *original* feature type used in a source gazetteer, and the *preferred terms* are the types defined by the research team as suitable for representing a *general* feature type classification [4]. Examples of the results sampled for 'school' are shown in Table 1.

Table 1 Feature Type List Compiled by the Alexandria Digital Library – Sample of "School"

/concept/#id	/concept/BT	/concept/descriptor	/concept/non-descriptor	/concept/NT	/concept/RT	/cor
9		administrative areas		school districts		
18			agricultural schools			
350	institutional sites	educational facilities				acad
350	institutional sites	educational facilities				agri
350	institutional sites	educational facilities				cam
350	institutional sites	educational facilities				colle
350	institutional sites	educational facilities				milit
350	institutional sites	educational facilities				scho
350	institutional sites	educational facilities				sem
350	institutional sites	educational facilities				trair
350	institutional sites	educational facilities				univ
350	institutional sites	educational facilities			library buildings	
350	institutional sites	educational facilities			research facilities	
679			military schools			
1005	administrative areas	school districts				
1006			Schools			

It is assumed within the ADL gazetteer that all source gazetteer feature types are recorded, where suitable, as 'non-preferred' and mapped to the 'preferred' category. Thus, the original descriptors are changed to suit the format of the ADL database.

By contrast, the popular geonames.org website has a feature type list which contains approximately 660 fine-level feature codes. The geonames.org gazetteer contains information from a host of source gazetteers, which unfortunately are never referenced directly in the metadata for each feature instance. Thus, the veracity of the data can never be interrogated at the original source. The geonames.org gazetteer utilises a list of feature codes, ranging from 2 to 4 letters, which aim at defining each feature in relation to its function. The feature codes are ranged in seven high-level categories, which are shown in Table 2. A sample of fine-level feature codes with 'school' in the description are displayed in Table 3.

Table 2 Geonames.org High-Level Feature Categories

Category CODE	Description
A	country, state, region,...
H	stream, lake, ...
L	parks, area, ...
P	city, village, ...
R	road, railroad
S	spot, building, farm
T	mountain, hill, rock,...
U	Undersea
V	forest, heath, ...

Table 3 Geonames.org Fine-Level Feature Types

Category CODE	Feature CODE	Feature Type	Description
S	CTRM	medical center	a complex of health care buildings including two or more of the following: hospital, medical school, clinic, pharmacy, doctor's offices, etc.
S	CTRR	religious center	a facility where more than one religious activity is carried out, e.g., retreat, school, monastery, worship
S	MSSN	mission	a place characterized by dwellings, school, church, hospital and other facilities operated by a religious group for the purpose of providing charitable services and to propagate religion
S	MSSNQ	abandoned mission	
S	NOV	novitiate	a religious house or school where novices are trained
S	SCH	school	building(s) where instruction in one or more branches of knowledge takes place
S	SCHA	agricultural school	a school with a curriculum focused on agriculture
S	SCHC	college	the grounds and buildings of an institution of higher learning
S	SCHL	language school	Language Schools & Institutions
S	SCHM	military school	a school at which military science forms the core of the curriculum
S	SCHN	maritime school	a school at which maritime sciences form the core of the curriculum
S	SCHT	technical school	post-secondary school with a specifically technical or vocational curriculum
S	UNIP	university prep school	University Preparation Schools & Institutions
S	UNIV	university	An institution for higher learning with teaching and research facilities constituting a graduate school and professional schools that award master's degrees and doctorates and an undergraduate division that awards bachelor's degrees.

There is scant information on geonames.org to analyse how the feature category lists were developed, and how the codes were applied. Further research is required into this area to gain a better appreciation of whether geonames.org and similar systems can be easily improved to match the best practices being developed by the UNSDI gazetteer framework.

As can be seen from the examples taken from ADL and geonames.org, even amongst the comprehensive nature of their feature mapping and categorisation tools, there are feature types unaligned between the two. For instance, geonames.org includes reference to *language schools* as a distinct feature type, whereas ADL has *school districts* as a feature type. These two samples from the datasets exemplify the issues associated with attempting to create comprehensive fine-level feature type catalogues. And within both these gazetteers it should be noted that the feature type categories are listed in English, with no known published translation to other languages. Further to this, within both ADL and geonames.org the school feature types are mapped rigidly to broader categories of 's' or 'educational facilities/administrative areas'. The potential for many-to-one relationships between the feature types and classifications is limited.

A classification method

An alternative perspective is to take a user-centric (or application-centric) perspective. For example, Open Street Map (OSM) applies symbolisation according to a set of "special" classifications whilst allowing others to be added. OSM is a user-generated, online map of the world. Updating and editing of the map relies on Volunteered Geographic Information (VGI) and the system for defining feature types, while having some established rules, is left mainly to the discretion of the contributor through a 'tagging' system. Feature types with map renderings are listed at http://wiki.openstreetmap.org/wiki/Map_Features, for example:

Table 4 Open Street Map Feature Tags

Education				
amenity	college		a College campus or buildings	
amenity	kindergarten		For children too young for a regular school (also known as playschool or nursery school).	
amenity	library		A public library (municipal, university ...) to borrow books from	
amenity	school		school and grounds	
amenity	university		a University campus	

Tags may be added for any purpose, but re-use of feature type tags supports map rendering. As can be seen in the example above, schools are categorised under the 'main elements' of *amenity* and *education*, and there are five main tag descriptions available for immediate use: *college*, *kindergarten*, *library*, *school*, and *university*. When compared to the catalogues used on ADL and geonames.org, the tags used by OSM seem rather limited. However, OSM encourage contributors to create their own tags, explaining that 'If you do not find a suitable tag in this list then feel free to make something suitable up as long as the tag values will be verifiable; over time you may find that the tag name is changed to fit will some wider consensus, however many good tags were used first and were documented later'. Thus, there are potentially more tags in use for educational facilities, but which are not published as part of the core OSM feature tags list.

Of interest for this discussion on feature type catalogues and categories, are the descriptors applied to feature instances within OSM. For schools, they are broadly tagged under the main element of 'amenity'. However, additional tags related to

the element of 'building' might also be suitable for most school-related features, and within the OSM platform multiple tags are catered for. Therefore, each feature instance might be tagged with multiple main elements and feature types. This approach reflects the multifaceted dynamics of human interaction with the landscape and our subsequent preference for having multiple categories for feature descriptions.

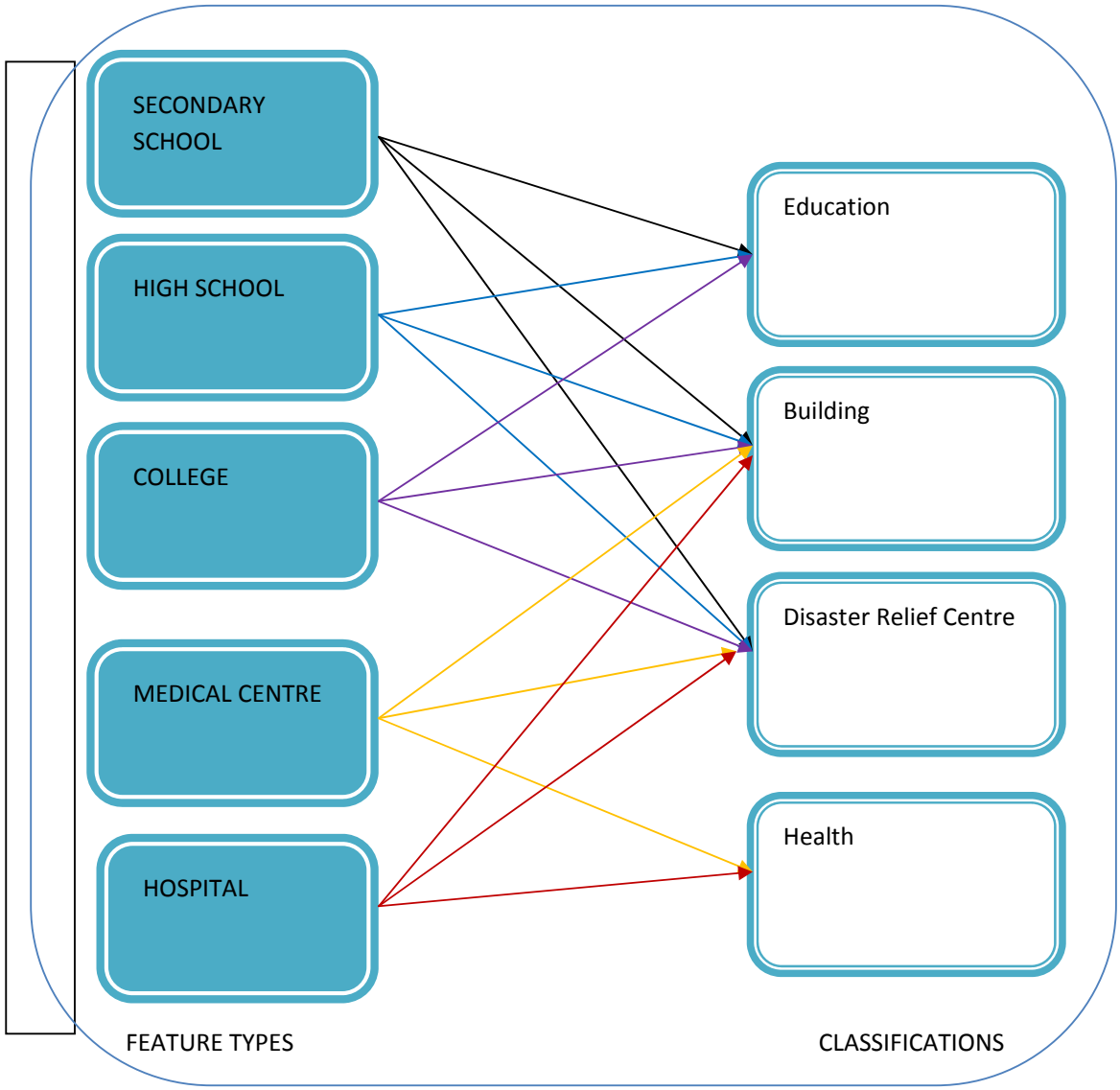
However, when dealing with government gazetteers developed by jurisdictional naming authorities, there is generally a need to have a more structured approach to feature type categorisation. While the OSM model of allowing for multiple 'tags' is appealing, there are drawbacks in regards to free-flowing categories being developed on an as-needs basis – particularly where there is potential for an exponential amount of feature types to be created for limited capacity gazetteers. Thus, the question remains- *How might UNGEGN best assist the international community when international gazetteer systems require feature mapping to be undertaken, yet there is a proliferation of online gazetteer feature types and classification systems?*

Our Proposal- A Common Feature Type & Classification CatalogueA user navigating a gazetteer (e.g. searching or browsing) is better served by a reasonably simple but comprehensive set of broad terms that are consistently applied. Different users may need different classifications – for example points of interest for a tourist is a significantly different application to emergency response, and may require a different set of classifications, but a similar level of broad scope for each term.

Research undertaken by Beard (2011) has indicated that there are demonstrable user-case scenarios associated with developing feature type classifications to which finer-level feature types are mapped. In particular, Beard [1] suggests that 'incomplete knowledge on the part of the user, semantic heterogeneity, and the ambiguity of natural language can create mismatches between user's feature type specification and the actual target. A user for example might search on *Otter Brook* while the intended target is *Otter Creek*. This type of search problem can be aided by lexical relationships of equivalence or similarity among feature types, relationships which are typically encoded by thesauri'

It is proposed that a common feature type list be approached through the development of one or more high-level classification schemes that can be implemented reliably in a multilingual, multi-script format. These would then become the target for detailed feature types in use in different source gazetteers. Additional levels of detail could be added as processes, interests and capabilities emerge to allow standardisation of a finer level of detail.

Table 5 High-Level Classification Model



The classifications should be developed through consultation with end-user groups, particularly those in the emergency response and humanitarian relief domains, who would be in a position to define their user-case scenarios for feature type categorisation. In all circumstances, the original source feature type should be preserved. This can be used to ensure integrity of the feature type catalogue and to enable the exploration of any finer grained classifications that may emerge over time.

Sources for the Common Feature Type Catalogue

Aside from the collection of feature type information from member states of UNGEGN, one source of fine-grained feature type classifications is the development of standardised international domain information models. International agencies operating under UN auspices should provide authoritative definitions that are referenced by other UN initiatives. Where they

publish formal information models and data specifications these can be used to source a well-designed set of target feature types, for example:

- **The International Hydrographic Organisation (IHO)** defines features used in maritime navigation
- **International Civil Aviation Organization (ICAO)**

These definitions are ideal candidates for common feature types in that they are already reconciled by domain experts and have been tested in practice through creation of data using the definition semantics.

Adoption of feature types defined by these types of organisation provides two practical advantages:

1. Ready-made governance processes for each subset
2. The ability to harvest data made available in these standardised formats directly into gazetteer collections.

Table 6 Example of Fine-Grain Feature Type Classifications from IHO S-100 standard

Feature Code	Description
SEAARE	Sea area/named water area
RAILWY	Railway
BRIDGE	Bridge
Hrbare	Harbour area (administrative
Hrbbsn	Harbour basin

It is evident that some of these feature types represent things that would be named in gazetteers, and whose official name is likely to be governed by IHO members, however in other cases names would be expected to be used from other sources. Thus "gazetteer" views of source data sets would be expected to filter out according to feature type. This provides both an opportunity and necessity to record the native feature type from such sources when establishing these governance rules for the subset to be included in, for example, national toponymic gazetteers.

Recommendations to UNGEGN

As noted in this paper, there are extensive issues to be resolved with regard to linking, comparing and aligning the multiple feature type lists available internationally. The UNSDI Gazetteer Framework Project research team proposes to UNGEGN members, the development of a Special Committee on Feature Types (perhaps as a sub-committee of the Working Group on Toponymic Data Files and Gazetteers, or the Working Group on Toponymic Terminology), which will bring benefit by feeding input from toponymic experts into the larger UNGIWG Spatial Data Infrastructure (SDI) program.

The Special Committee would need to have Terms of Reference established, which the research team would be pleased to assist with, and they could be tasked with determining the optimal methods for developing either end-user-defined classification schemes or extensive fine-level feature type lists mapped to jurisdictional gazetteers. The other proposed benefits of the development of a feature type and classifications list are extensive, and include the ability for naming authorities to identify areas for increasing the scope of their gazetteer data collection methods (as discussed in the *Four Faces of Toponymic Gazetteers* paper submitted by Australia for this conference).

References

1. Beard, K., *Organising Relationships to Aid Place Name Searches.* Journal of Spatial Information Science, 2011.
2. Godoy, J., J. Atkinson, and A. Rodriguez, *Geo-referencing with semi-automatic gazetteer expansion using lexico-syntactical patterns and co-reference analysis.* International Journal of Geographical Information Science, 2011. **25**(1): p. 149-170.
3. Martins, B., *A supervised machine learning approach for duplicate detection over gazetteer records.* Lecture Notes in Computer Science- Geospatial Semantics, 2011. **6631**: p. 34-51.
4. Hill, L., *Core Elements of Digital Gazetteers: Placenames, Categories and Footprints.* Lecture Notes in Computer Science- Research and Advanced Technology for Digital Libraries, 2000. **1923**: p. 280-290.
5. KeBler, C., et al., *Bottom-up Gazetteers: Learning from the Implicit Semantics of Geotags.* Lecture Notes in Computer Science- Geospatial Semantics, 2009. **5892**: p. 83-102.
6. Hill, L., et al., *Alexandria Digital Library: User Evaluation Studies and System Design.* Journal of the American Society for Information Science and Technology, 2000. **51**(2): p. 246-259.
7. Baglioni, M., et al., *Building Geospatial Ontologies from Geographical Databases.* Lecture Notes in Computer Science- Geospatial Semantics, 2007. **4853**: p. 195-209.
8. Tanasescu, V., *Spatial Semantics in Difference Spaces.* lecture Notes in Computer Science- Spatial Information Theory, 2007. **4736**: p. 96-115.