Notes from the 7th March 2007 Meeting of the UNGEGN Working Group on Toponymic Data Files & Gazetteers*
[with draft toponymic data exchange standard]

A joint session of the UNGEGN Working Group on Toponymic Data Files and Gazetteers and the EuroGeoNames Reference Group and Consortium was held at the Palacio de Zurbano, Madrid, Spain, 5th - 7th March 2007. The Working Group session was attended by sixteen members from ten countries.

The Working Group discussed its website. Mr Leskinen offered to re-create it and to take on maintenance responsibilities. Suggestions for content included a draft exchange standard, the 1998 Working Group on Toponymic Data Exchange Formats and Standards Report as updated for the UN Technical reference manual, relevant papers from previous UNGEGN sessions, exemplar data models, and links to national geographical names databases.

Liaison with the Unicode Consortium was considered by the Working Group to be important, in order to communicate anomalies in Unicode standard to the Consortium and to collaborate to resolve these. The co-ordinators would be Mrs Burgess, Mr Päll and Mr Viechnicki. A representative from the Unicode Consortium would also be invited to the 9th UN Conference on the Standardization of Geographical Names (UNCSGN).

UNGEGN had established liaison with the International Organization for Standardization (ISO) Technical Committee 211 (Geographic Information/Geomatics) in 2002. The Working Group agreed that a representative of TC 211 should be invited to the 9th UNCSGN. The Convenor would resume the co-ordination process through contact with appropriate individuals at the US National Center for Geospatial Information Standards.

The Working Group would update and expand its draft toponymic data exchange standard (see draft below). Once the draft had been examined by Working Group and then UNGEGN members, it would be presented to ISO with a view to its being adopted as a formal ISO standard for toponymic data exchange. The Convenor would co-ordinate the review of the draft.

The Working Group agreed that it would focus on the provision of information in the form of the development of standards, rather than attempt to become closely involved in practical programmes and/or workshops of toponymic data exchange. With this in mind, the Working Group's terms of reference would need to be reviewed and amended.

Mrs Burgess gave a presentation (to be posted on the Working Group's website) on the Working Group's contribution to the UN Technical reference manual which was about to be published. Section II comprised an update of the Working Group on Toponymic Data Exchange Formats and Standards Report. The annexes to this report contained a detailed study of encoding standards treating Roman-script characters used in geographical names, along with a draft toponymic data exchange standard (see above). The Working Group agreed that these annexes should be maintained and the original authors of the report (specifically Mrs Burgess and Mr Päll) would take on that responsibility, posting updates on the Working Group's website. The Working Group considered the expansion of document to cover non-Roman writing systems. Although an enormous task, the Working Group would investigate the feasibility[1].

---

* Compiled by Mrs. Caroline Burgess, United Kingdom.

[1] It was later mooted that other Working Group/regional UNGEGN Division assistance could be sought.

The UNGEGN Chair gave a presentation on the status of the UN Geographical Names Database. UNGEGN Divisions would be approached to provide input and verification of data. The Working Group reviewed the presentation and discussed information integrity. A Resolution of the 9th UNCSGN would be drafted to support the continuing evolution of the database. The population data being included could be found on the UN Statistics Division page at: http://unstats.un.org/unsd/demographic/sconcerns/densurb/urban.aspx. A copy of the UN database model would be posted on the Working Group website.

The Working Group recognised that there were differences in capabilities supporting the development of toponymic data files and would consider whether the Working Group could assist in narrowing this capability gap.

The importance of training was recognised by the Working Group and co-ordination with the Working Group on Training Courses in Toponymy would be encouraged to improve the content of courses in relation to toponymic data files and the basic elements of database construction. The Working Group also considered the collation of previous UNGEGN working papers into a single handbook to aid organisations in development of toponymic databases.

The Working Group received copies of a new UN brochure, compiled by the Working Group on Publicity and Funding, for review and comment. The Working Group recommended that a Resolution be drafted for the 9th UNCSGN to support the continuation of an UNGEGN publication programme.

The UNGEGN Chair reported recent communication with representatives of GoogleEarth. A number of geographical names issues had come to light with GoogleEarth's use of third party data. A pilot project had been initiated with Finland to seek methods of improving the quality of the names data. The Working Group acknowledged the need for UNGEGN liaison. GoogleEarth had made a verbal commitment to attend the 9th UNCSGN.

Mr Ferland gave a presentation to facilitate discussion among the Working Group on the future of gazetteers and gazetteer services. The Working Group agreed to set up a sub-committee to consider such issues further. Mr Ferland's presentation would be posted on the Working Group's website.

The Convenor reported on a Digital Gazetteer Research and Practice Workshop, which had taken place in Santa Barbara, CA, USA in December 2006. The University of California website at www.ncgia.ucsb.edu contained a draft report of the meeting and a research agenda.

The Convenor also outlined an US NGA experiment to host its Geographic Names Data Base in a WikiEnvironment to encourage collaborative toponymic information exchange. It was agreed that an entry on the Working Group should be prepared for posting on the Wikipedia site.

The Working Group discussed preparations for the 9th UNCSGN to be held in New York in August 2007. Two poster-format panels for an exhibit area at the Conference would be contributed by the Working Group. Members would pass ideas and recommendations to the Convenor who would arrange the creation of the posters.

The Working Group considered it would be beneficial for a number of presentations to be given at the 9th UNCSGN, specifically by GoogleEarth, the Unicode Consortium, ISO/TC211, the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), the UN Geospatial Information Group (UNGIWG), Michael Goodchild of the US National Center for Geographic

Information and Analysis (NCGIA) and the University of California Santa Barbara, and Allen Carroll of US National Geographic Society. An invitation to the relevant ISO technical committee concerning language standards should also be considered. The Working Group also recommended that small workshops be held during the Conference, on data exchange standards, legal issues relative to place names (eg place names acts), and WikiGazetteers.

The next meeting of the Working Group would be held during the 9th UNCSGN. Future joint meetings with other UNGEGN working groups/divisions/EuroGeoNames would also be investigated.

8th March 2007

Draft toponymic data exchange standard

Part 1 Data elements

The following data elements, identified in Resolution 4 of the first United Nations Conference on the Standardization of Geographical Names, comprise the minimum set of critical geographic names information for digital exchange.

1.0 GEOGRAPHIC NAME– the spelling of a standardized name referring to a feature. Spellings of variant names, if any, follow the standardized name with suitable delimiters. An explanatory note accompanying the data set must define all conventions used in the presentation of GEOGRAPHIC NAME; e.g., sorting order, reversal of generic terms, meaning of special flags such as asterisks, etc.

Data Type: text. An explanatory note accompanying the data set must define the text format and encoding; e.g.,
char, ASCII
char, ISO 8859 (2)
char, KOI 8
wchar_t, ISO 10646

2.0 FEATURE TYPE – a characterization of the kind of feature represented by GEOGRAPHIC NAME. An explanatory note accompanying the data set must define the characterization scheme employed.

Data Type: text

3.0 LOCATION – the position associated with GEOGRAPHIC NAME or FEATURE TYPE expressed in latitude and longitude (based on the Prime Meridian, Greenwich). For most toponymic data sets, the preferred format would be positions specified by degrees and minutes, rounded down to the nearest integer minute. This format allows future update to include seconds or decimal minutes. An explanatory note accompanying the data set must identify alternate formats if applicable; e.g., decimal degrees, integer degrees and minutes, integer degrees and decimal minutes, etc.

3.1 LATITUDE – the latitude value of LOCATION.
Data Type: integer
Domain: 90 00S <= LATITUDE <= 90 00N

3.2 LONGITUDE – the longitude value of LOCATION.
Data Type: integer
Domain: 180 00E <= LONGITUDE <= 180 00W

4.0 ADMINISTRATIVE AREA – an identification of the country and administrative division in which LOCATION falls.

Data Type: text
Domain: (the domain is the content of ISO/DIS 3166, Parts 1 and 2)

5.0 MAP SHEET REFERENCE – an identification of the sheet (within a standard national map series) containing LOCATION. An explanatory note accompanying the data set must provide citation for the referenced map series.

Data Type: text

Part 2 Metadata[2]

The notes to Part 1 of this annex refer to additional information required by the user of a data set to interpret competently the data contained therein. This additional information is generally referred to as *metadata*.

Metadata serves to answer four questions:

- What sets of data exist for a geographic location?
- Does a set of data meet a specific need?
- How is a set of data acquired?
- What information is required to process and use a set of data?

Metadata characterizing geographic names data sets would typically comprise the following subsets:

- Identification information – basic information about the data, including publisher.
- Text encoding standard – identification of the national, international, or proprietary standard used to represent the text digitally in the data set.
- Data quality information – a general assessment of the quality (accuracy, currency) of the data set.
- Spatial data organization (if applicable) – the mechanism (text, point, vector, raster) used to represent spatial data in the data set.
- Spatial reference information (if applicable) – the description of the reference frame for coordinates in the data set.
- Entity and attribute information – the description of the content of a data set including entity types, their attributes and domains (Part 1 of this Annex qualifies as entity and attribute metadata).
- Software information – identification of the software configuration used to create the data set, including identification of operating system and application.
- Distribution information – information about the distribution of and options for obtaining the data set.

Part 15 of ISO 15046 is a draft standard for metadata pertaining to geographic information. It is based largely on the Metadata Standard developed by the United States Federal Geographic Data Committee (FGDC). While it was beyond the scope of the UNGEGN Working Group on Toponymic Data Exchange Formats and Standards to define rigorously a metadata standard specific to toponymic data, it is appropriate and necessary to emphasize the importance of describing the content of geographic names data sets in sufficient detail to ensure their utility in a program of exchange of digital information.

Part 3 Illustration

The following listing is an abbreviated example of a print-out of a digital toponymic data set adhering to the guidelines recommended by the Working Group. Two files are represented; the first is the toponymic data, the second is a text file (conventionally known as a “readme” file)

[2] The information presented here is summarized from Content Standard for Digital Geospatial Metadata, Federal Geographic Data Committee (USA), April, 1997

containing the metadata necessary for a recipient to determine the data's content and utility. These examples are provided for the sole purpose of illustration.

Data.txt

Aakre Peatus	RSTP	5804N	2616E	EE	19	NO35-05
Aaksi Saar → Aksi	ISL	5935N	2505E	EE	01	NO35-01
Aamse	PPLX	5855N	2342E	EE	07	NO34-06
Aamsi → Aamse	PPLX	5855N	2342E	EE	07	NO34-06
Aandu	PPL	5909N	2443E	EE	13	NO35-01
Aardla	PPL	5818N	2646E	EE	18	NO35-05

Readme.txt

This file describes the content, lineage, and utility of the toponymic data contained in the file data.txt.

Data Identification: Place name spellings for features in Estonia.

Text encoding standard: Eight-bit ASCII (ISO 8859 Latin 1). Two characters (Roman small letter c with caron; Roman small letter i with macron) are encoded in a proprietary scheme, and may be viewed using the Monotype Special G1 family of fonts.

Data quality: Name spellings are current for most features as of 1993, and have been taken from the following authoritative Estonian sources:

Eesti Topograafiline Kaart, 1:200,000, 1993.
Eesti Ülevaatekaart, 1:400,000, 1993.
Eesti Sood, 1:400,000, 1993.
Eesti Veed, 1:400,000, 1991.

Settlement names were taken from a list provided by Eesti Keele Instituut, Tallinn, in 1994.

Data organization: Data is provided alphabetically by name spelling. Data elements within a record are tab delimited (ASCII character 09). Records are delimited by carriage return (ASCII character 0D) Single geographic coordinates are provided for each record.

Data reference: Coordinates are referenced to the World Geodetic System 1984.

Entity and attribute information: Each record contains the following data elements and associated attributes:

Geographic name. Name spellings conform to standard Estonian orthography with the exception that initial letters of generic terms are capitalized. Variant names are cross referenced to standard names by use of the greater than sign (ASCII character 3E).

Feature type. A five character designation denoting the type of feature to which the name refers. The designation scheme is that employed by the U.S. Board on Geographic Names in its data base of foreign geographic names. This data element is left justified, and space filled (ASCII character 20) in cases where the designation is fewer than five characters.

Latitude. The geographic latitude of the named feature, taken at the centre of spot and areal features, and at one end of linear features. The data element is five characters in

length; the first two represent integer degrees, the second two integer minutes (rounded down), and the fifth the compass point N.

Longitude. The geographic longitude of the named feature, taken at the centre of spot and areal features, and at one end of linear features. The data element is five characters in length; the first two represent integer degrees, the second two integer minutes (rounded down), and the fifth the compass point E.

Country code. The country code digraph taken from ISO 3166 "Codes for the representation of names of countries and their subdivisions."

Administrative division code. The two character administrative division code taken from ISO 3166 "Codes for the representation of names of countries and their subdivisions." The administrative division identified for each name is the division in which the geographic location for the record falls. Features that cross administrative or international boundaries are given the general code 00.

Map sheet reference. The map sheet number of the standard 1:250,000 map series on which the geographic coordinate of the record falls. The feature and its name may not necessarily actually appear on the referenced sheet.

Software requirements. This data set was compiled using Microsoft Word 7.0 for Windows 95. The only known special requirement is the use of the Monotype Special G1 font for display of certain characters.

Distribution information. There are no restrictions on the distribution and use of the information contained in these data sets. Additional copies of this data set may be obtained from the producer.

Producer point of contact. This data set was produced by the staff of the Foreign Names Committee of the U.S. Board on Geographic Names. Questions, corrections, and requests for additional information may be referred to:

Executive Secretary for Foreign Names
U.S. Board on Geographic Names
National Geospatial-Intelligence Agency
Political Geography Division (PRP)
4600 Sangamore Road
Bethesda, Maryland 20816-5003
USA