United Nations E/CONF.98/18/Add.1

Economic and Social Council

Distr.: General

25 June 2007

Original: English

Ninth United Nations Conference on the Standardization of Geographical Names
New York, 21 - 30 August 2007
Item 8 of the provisional agenda*
Economic and social benefits of the national and international standardization of geographical names.

User/Business requirements of a European Geographical Names Infrastructure and Service

Submitted by the Netherlands**

* E/CONF.98/1

** Prepared by Ferjan Ormeling (the Netherlands), Leader EuroGeoNames Workpackage "User/business requirements".

USER/BUSINESS REQUIREMENTS OF A EUROPEAN GEOGRAPHICAL NAMES INFRASTRUCTURE AND SERVICE

Ferjan Ormeling
Leader EuroGeoNames Workpackage "User/business requirements"
Utrecht University
f.ormeling@geo.uu.nl

Introduction

In 2005 a research consortium consisting of the Bundesamt für Kartographie und Geodäsie (Germany), Utrecht University (the Netherlands), the Land Survey Offices of Austria and Slovenia, EuroGeographics (the association of European national mapping agencies) and ESRI Germany did a survey/inventory amongst European national mapping agencies in order to assess the availability, quality, accessibility and responsibilities regarding geographical names data in Europe.

This survey was targeted at finding out the feasibility to create a Europe-wide virtual geographical names service, based on national geographical names servers, controlled and kept up to date on a national level. A questionnaire was sent out early 2005 and was answered by 24 countries; its results were analysed and evaluated during the summer and a preliminary report was produced September 2005, checked by the respondents in October and finalised in November (EGN 2005a).

The questionnaire results showed a widely deviating array of practices and procedures, and an overwhelming majority was in favour of cooperating with the realization of a European geographical names service, to improve the availability and usability of standardised official geographical names data, harmonization of the European geographical names digital data by developing common specifications and a joint data model. The aim to implement a European geographical names data infrastructure under the umbrella of EuroGeographics, was integrating the national data sources. The resulting network will be kept up to date in a consistent way and maintained at source level by the responsible proprietary organisations.

The main names categories discerned by national mapping agencies are largely consistent, in the sense that a broad subdivision into names of populated places, hydrographic features, administrative units, terrain features and geographical areas is valid everywhere; almost all countries could provide names taken from maps at scales 1:50 000 or larger. Gazetteers were available for most of Western Europe, based on 1:50 000 map series, updating occurred mostly on a yearly basis at least for names of administrative divisions and populated places. The geographical names were standardised (see figure 1) by either national mapping agencies, national statistic offices, geographical names boards or other governmental institutions. Many European countries (17 out of the 24 that answered) already have a names information service, although the metadata the names files are tagged with are still highly divergent. The number of records contained in the national names files varies widely; 12 countries had direct links between the names databases and topographic data. Names data provision was possible through web servers in 11 out of 24 countries. Most name servers could provide sorted extract

output. The attribute information linked to the geographical names varied greatly (see table 1) as did the precision for point features in the attribute information.

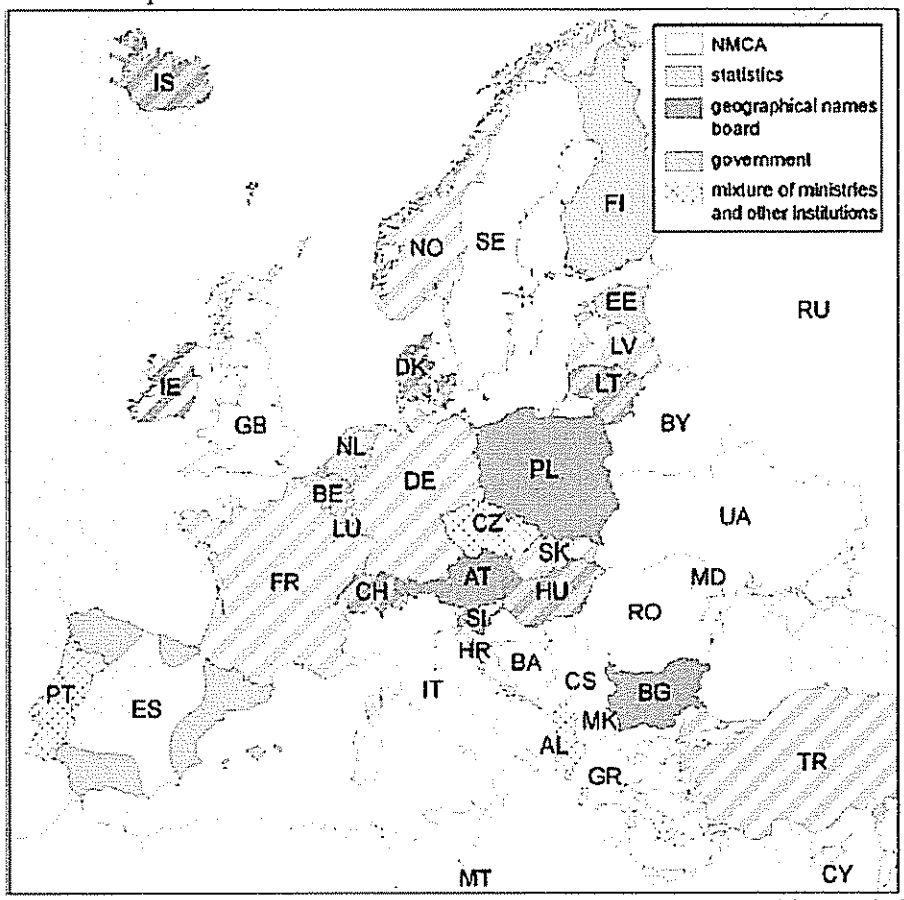


Figure 1 Dominant institutions approving geographical names in Europe (white: no information)

For the names databases, database packages in use consisted of various versions of Oracle, Ms Access or SGL. The names could be made available in many formats, and the languages supported by the operating systems also varied widely. Most countries had lists of exonyms available.

After this inventory, an application for the main phase of the EuroGeoNames project (to develop a European geographical names infrastructure and services) was sent in to the EU Inspire for its 2006 eContent plus application round (EGN2005b), and this was accepted; the project started September 2006 and is supposed to be completed by February 2009; for this application the consortium had been extended to include as well Edinburgh University (UK), GeoDan (Netherlands) and GeoTask (Switzerland).

Attribute information linked to the names in the various national databases	Feature coordinates	Name placement coordinates	Feature category	Feature object ID	Map scale indicators	Statistical classification	Name status	Language	Pronunciation	History	Other	Height	Number of inhabitants	Map sheet number	Language status	Size + style character	Gender	Name sources	Variant names
Albania																			
Austria	X		X	X								X							
Belgium																			
Bulgaria	X	X	X	X	X		X	X	X										
Croatia		X	X	X															
Cyprus	X	X		X								X	X						
Czech Republic		X	X	X			X												
Denmark	X	X	X	X	X	X	X						X						
Estonia	X	X	X	X		X	X	X		X									
Finland	X	X	X	X	X	X		X						X	X	X			
France		X	X		X														
Germany AdV	X		X	X								X	X				X		
Germany BKG	X		X	X		X											X		
Germany StAGN	X		X			X													
Great Britain	X	X		X	X	X	X	X	X	X									
Hungary	X		X	X			X					X	X					X	X
Iceland		X	X	X						X						X			
Ireland		X				X													
Latvia	X		X	X	X	X	X		X				X			X			
Lithuania		X	X	X															
Netherlands	X	X	X				X												
Norway	X		X	X	X	X	X	X										X	
Poland	X		X	X	X	X	X						X						
Portugal		X	X																
Slovakia	X		X	X		X	X					X							
Slovenia		X	X	X	X			X											
Spain	X		X		X	X	X	X											
Sweden		X	X	X		X	X	X											
Switzerland	X		X	X	X														
Turkey	X		X	X	X	X						X	X						

Table 1: Attribute information linked to the names in the various national databases/geographical names repositories

Contents of the EurogeoNames main phase project

The general objectives for the EGN infrastructure and services are:
-aggregation of European public sector geographical names information in order to provide harmonized access to a multilingual pan-European data infrastructure for the citizen, governance and value-added services
-increased availability and accessibility of authoritative national geographical names data
-increased use or geographical names in spatially-related decision making processes
-increased re-use and value-adding by commercial enterprises
-stimulation of European national mapping and cadastral agencies towards better integration of geographical names data into national SDIs.

Specific objectives for the EGN infrastructure and services are:
-support of all officially recognised minority languages
-development of a network of geographical names experts

-easy and rapid linkage of exonyms with their corresponding endonyms
-attainment of cost efficiencies in the collection, handling, storing, maintenance and distribution of geographical names data
-development of an implementation plan for continuing and extending the services beyond the end of the project (EGN 2005b)

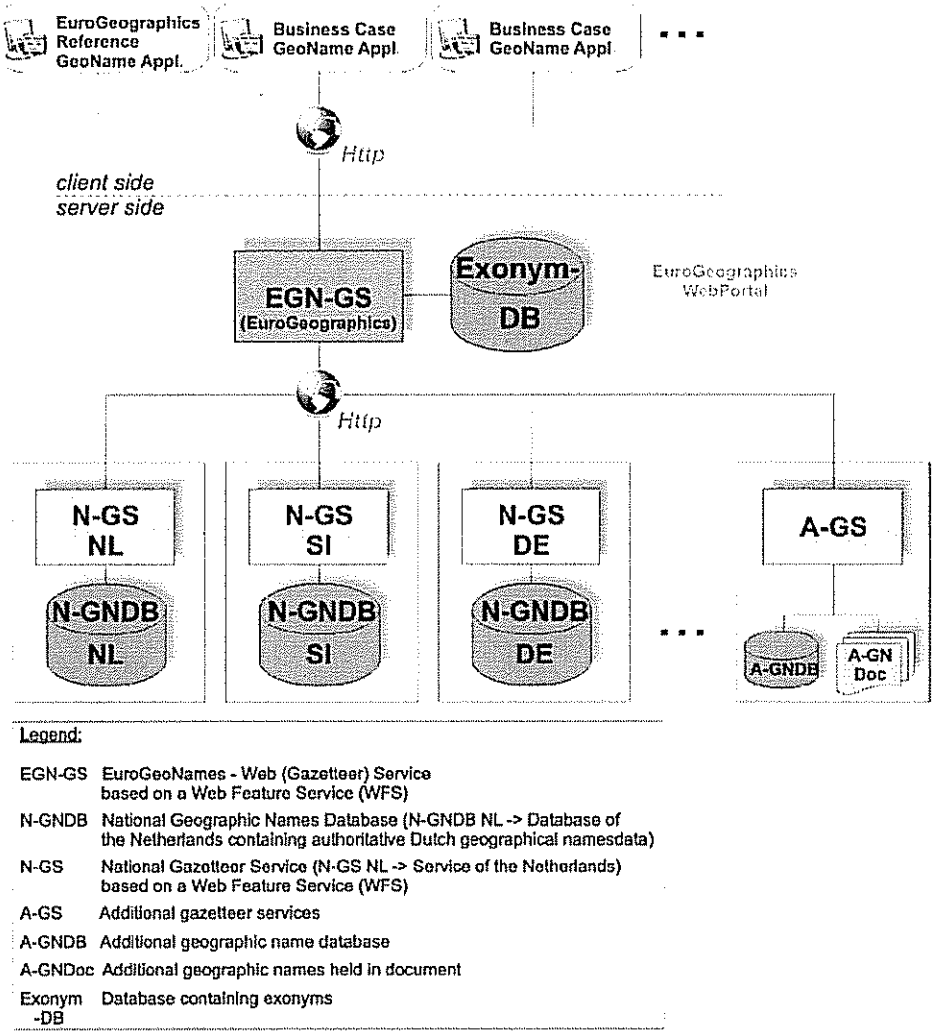


Figure 2 Possible service architecture for a European geographical names service

Figure 2 shows a possible service architecture for a geographical names service as planned in the EuroGeoNames project, used for a geographical names reference application hosted by EuroGeographics or BKG and some business cases. In order to realise this, the project main phase has been subdivided into a number of work packages:

Work package 1 Project coordination and financial management
Work package 2 User/business requirements
Work package 3 Data selection
Work package 4 Data model definition
Work package 5 Exonym management

Work package 6 EGN webserver infrastructure Work package 7 EGN Web GIS reference application Work package 8 EGN Web GIS application in a commercial context Work package 9 Test environment harness Work package 10 Business model and marketing concept Work package 11 Implementation plan Work package 12 Awareness and dissemination Work package 13 Assessment and evaluation

Box 1 division of the EGN main phase project into work packages

Objectives and tasks of EGN Work package 2

The objectives of EGN Work package 2 are to analyse the market in terms of potential applications and related business actors and stakeholders from both public and private sectors and to provide a detailed description of user and business requirements based on the outcome of the market analyses. This description will form the basis for the development of the data model (WP4) and associated services (WP6-8, 10).

These objectives can be reached by fulfilling the following tasks:
Task 2.1 **Prepare Draft and Final Reports on User/Business requirements**. Describe the most common user groups and their needs and wants (including categorisation of user profiles, e.g. within public and private organisations, regional, national and European perspectives, etc). Investigate and compile existing surveys on user requirements as well as existing analyses on business issues and market-oriented requirements for reference data and associated services. Prepare a draft report before the workshop (task 2.3). Prepare a final report after the workshop to be mandated by the Consortium.
Task 2.2 **Potential application definition**. Compile and describe potential commercial applications and uses of an EGN web service before the workshop (task 2.3)
Task 2.3 **Organise a workshop on user/business requirements**. In this workshop the contents of the draft report on user/business requirements have to be discussed, and it has to be determined what core geographical names are needed on the European level (data content, data structure, coverage, level of accuracy, data transfer and information access services, handling of updates, management services among the partners of the information system). In the workshop the user/business requirements (resulting in a preliminary information model) need to be checked.

The major result of the working package 2 would be an information model that can be used as a basis for the data model, the EGN Web service development, the design of the applications and the business model. But the report of the WP2 should also contribute to the formulation of the vision and strategies in WP10. It will contribute towards formulating the vision, the strategy and the positioning. The rest of this paper contains a summary of the report of WP2, divided over the sections Data sources, Geonames functions and user groups, Market analysis and inventory of current commercial applications, Current offer of geonames web servers and comparison with the EGN proposal, and the information model. For the full report, see EGN2007a.

Data sourcing

Presently, geonames files are very much based on the paper maps they were taken from, as can also be seen from the fact that in some countries geoname coordinates still refer to the locations of the names on the map instead of the locations of the named objects. In future, this will be reversed and the names on the map or in gazetteers will be a selection of the national names databases or the national topographic databases, with a tendency towards the latter, where the names will be attributes of topographic objects.

Geonames are decided on by (local) authorities. Topographical fieldwork is considered the best means to have geonames information officially collected, both in the field and from local government, as this should be free from bias or commercial considerations. This state monopoly is increasingly under pressure, as mapping activities are being outsourced and there is an increased investment of commercial geospatial information suppliers in fieldwork activities, such as for yellow pages, route planning systems and tourist information systems. In future some form of symbiosis between public and private geodata collecting systems might be envisaged.

The rate of change in geographical names is considerable. It is supposed to be in between 5 and 23% per revision cycle for geonames (and between 5 and 15% for street names); the higher percentages would refer to changes in mapping procedures, changes in orthographic rules or in procedures regarding minority names.

Geonames functions and user groups

The following use groups and functions have been discerned, that were reflected in a number of use cases described (see table 2) in the report. The function of these use cases was to find a number of different cases to base the information model on and to derive the required functionality from.

functions/ use groups	normalisation = look up	translation	indexing	geocoding ~ geoparse	geo-indexing	reverse look up
finance, insurance	xxx	xx				
web sales, tourism	xxx 1	xxx 1	xxx	xx	xx 1	
marketing, health	xxx 9	x 9		xxx 9	xxx	
media	xxx 8	x 8	xx			
distribution	xx	xxx	xx	x	xx	
spatial planning	xx 7	xx		x 7	x	
map data production	xxx	xxx	xx	x	x	xx

emergency services, lbs	xxx	x	x 4	xxx 4	x	xxx 11
science	xxx 5	xxx 5		xxx 5	x 6	x 6
individual users	xxx 2, 3	xxx 2, 3	x	xxx		

Table 2 - Functions for the use groups discerned, on the basis of needs (xxx), wants (xx), nice to have (x). Numbers refer to use cases (highlighted).

On the basis of the functions to be played by the names service for the various user groups or fields discerned, the following use cases were described:
1) hotel booking use case,
2) EGN name server use case,
3) geoportal map application use case,
4) emergency map use case,
5) geoparsing use case,
6) historical research use case,
7) metadata search use case,
8) geonames checker use case,
9) real estate use case,
11) coordinate emergency use case,

To get an idea of the description method used to describe these use cases, in Appendix 1 a description on a geoparsing use case.

From the use cases, the following requirements for the EGN names service emerged:

Content: all place names and their variant names in all European languages, their gender, number and their coordinates.
-All major lakes and rivers, mountains, physical and administrative areas, all named landscape features. Named areas should have polygons. Previous, replaced names should still be in the database
-Finance and Insurance and Use case 4 and 9 also require street addresses and major points of interest
-Use case 8 also requires pronunciation information
Functionality: at place name input show coordinates or bounding boxes (and show adjacent names), show feature types, show variant names, fuzzy name search (also: sounds as:); show name on a map (use case 2), combined query (feature and name or feature and area or feature and coordinates)
Use case 8 requires a pronunciation facility. Proximity analysis for cases 1 and 6
User interface: ease of use is crucial. Purpose of UI elements must be immediately clear. Multilanguage
Integration: available 24/7, protocol agnostic but ADL or OGC preferred. Ability to perform multiple iterative searches
Data quality: names data should be annually updated. Spatial accuracy from 10 (case 4) up to 100m (case 3).

Box 2: EGN geonames service requirements phase 1

Market analysis and inventory of current commercial applications

A study of the geospatial information marketing possibilities identified unexpected market players such as insurance companies and health research. Furthermore, commercial users wanted to use wide-spread software technology, they wanted straightforward incorporation of the gazetteer service in their existing infrastructure and they wanted an easy exchange of information, both through import/export and through open service capabilities.
In the inventory, it was registered whether operators were from the public or private sector, their business model and platform was indicated, service providers and search categories used (place names (9), administrative units (7), addresses (4), postal codes (3), motorways (2), other areas (2).
Most common search categories in current commercial applications were complete address information and administrative area searches. The accuracy required was point coordinates for address information and bounding boxes for all other name categories. Annual updating of information seemed to be the norm.

Current offer of geonames web servers and comparison with the EGN proposal

A survey of current free commercial geonames servers on the web shows 7 providers (Geonames, GEOnet Names Server (GNS), Getty thesaurus of geographical names, the Fuzzy Gazetteer, the ADL Gazetteer service, Earth Search and World gazetteer. Apart from FuzzyG all of them use the names taken from the GNS, maintained by the US National Geospatial Intelligence Agency (NGA) and the US Board on Geographical Names. The Fuzzy Gazetteer was maintained by the European Commission and the European Joint Research Centre. Attribute data all these servers had in common are names, coordinates, feature codes, country codes and functionality they all shared was the location of the named objects on the map. Half of them worked with unique name or object IDs.

The unique selling points (usp) of the EGN names service for Europe are, that the names data provided are from a primary source that is continuously updated and more detailed than GNS data. It is closer to the experts that collect the names, there is better quality control through official cooperation and it is based on European standards. The current geonames webservers reviewed score as follows in comparison with EGN regarding these usp's: (see table 3)

EGN usp's	**EGN**	**current geoname servers**	**current route planners**
primary data	yes	no	?
official data	yes	?	partly
high quality data**	yes	?	yes
up-to-date data	yes	no	yes
complete coverage	not yet*	no	depends on category

according to European standards	yes	?	yes
including street level data	no	no	yes
coverage	Europe	whole world	Europe +

Table 3 - Comparison between EGN geonames service, current name servers and route planners. *) complete coverage is envisaged. **) comparable to 1:50 000, coordinates to 10m

Information model

Based on the requirements in box 2 (taken from the use cases) 5 main name categories are discerned (one of which, the man-made objects category, is further subdivided), and pronunciation is added as a name attribute. Based on the user interface requirements in box 2, language should be an attribute, as well as the date a name was last changed. Street addresses and postal codes came up as a requirement in two use cases as well as in the market analysis and in the commercial applications. In view of the fact that they were beyond the scope of EGN, and also because all of the competing geoname servers did not provide street addresses and postal codes either, it was decided not to require street names/addresses for the time being. It should be discussed whether to leave room for extending the EGN functionality with street address matching capabilities in future, so that EGN would be able to compete with this address matching possibility in Route planners. In view of the fact that after street address matching the second most frequently used search category was administrative area search, it was decided to include hierarchical objects in the information model.

The most extensive competing geoname servers all had ID's, either for the topographical objects named or for the geographical names. They also had information on the status of the names and the dates this status changed. As the official character of EGN is stressed, information on this aspect of the names should be made available. As it was decided not to go into data enrichment and leave that to value-added resellers, the present attributes are rather few. The metadata for the topographical objects would refer to data quality aspects such as up-to-dateness, completeness and geometrical accuracy.

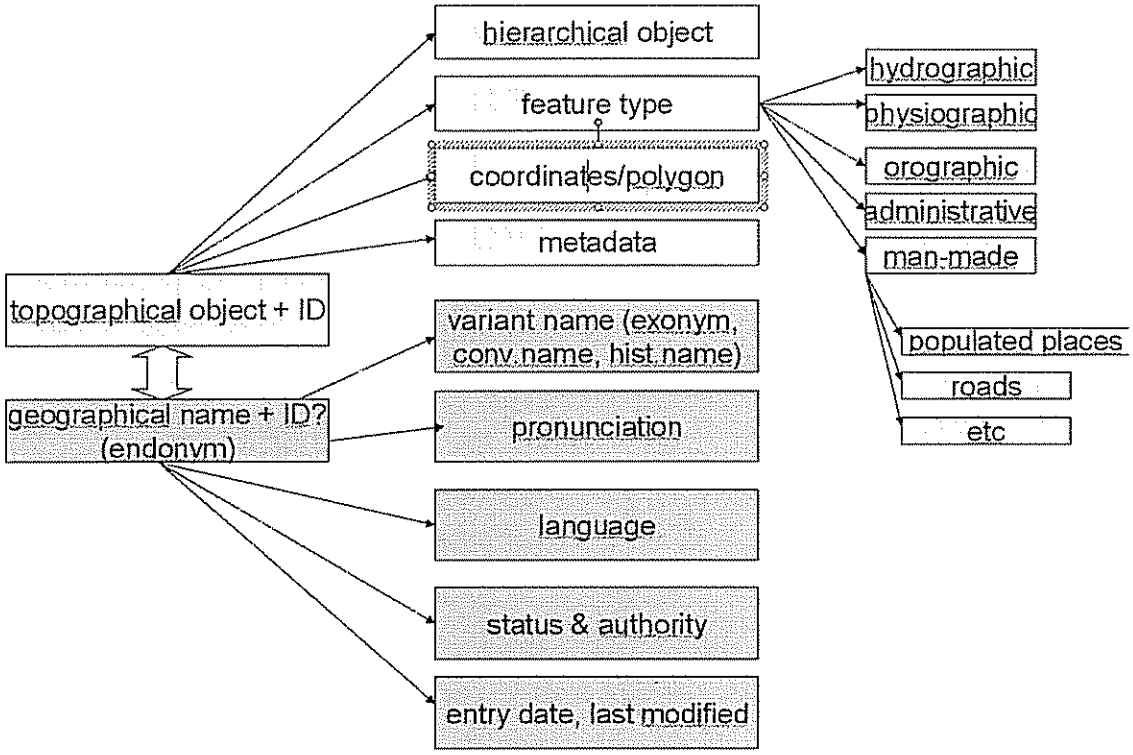


Figure 3 Proposed EGN Information model, based on use cases and market analysis

Literature

EGN 2005a - SI-EGN Final Report. Survey/inventory on the state of the art of European geographical names data sources and assessment of a future European geographical names data infrastructure (SI-EGN). Edited by Utrecht University and Federal Agency for Cartography and Geodesy, Frankfurt 2005

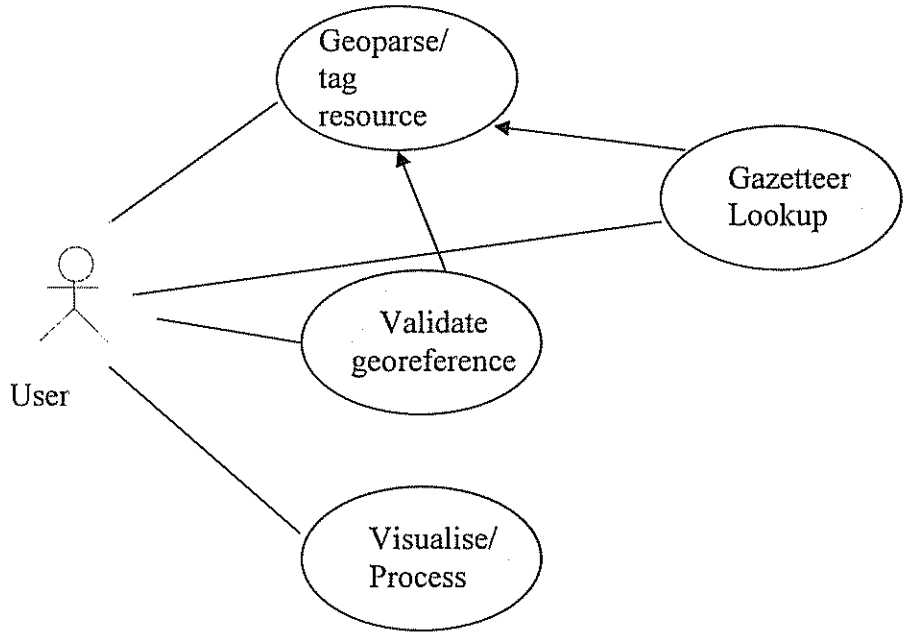
EGN 2005b - EGN/EuroGeoNames-developing a European geographical names infrastructure and services. Description of work. Annex 1 of EGN project application. Edited by Federal Agency for Cartography and Geodesy, Frankfurt 2005

EGN 2007a - Report EGN WP2 User/business requirements. Edited by Utrecht University, ESRI Germany, GeoDan and GeoTask). Utrecht 2007

- EGN 2007b: Report EGN Workpackage 3 Tables of quantity. Prepared by the Surveying and Mapping Authority of the Republic of Slovenia. Ljubljana 2007 (unpublished).

Appendix 1 Geoparsing Use Case (produced by Edina)

Part 1: UML use case diagram for Geoparsing.



Part 2: Narrative explanation.

The GeoParser demonstrator use case described here is a tool developed by EDINA that allows users to upload web pages, text files, metadata records, xml etc , which can then be parsed for geographical names. These are then checked against a gazetteer to obtain explicit geographical coordinates for the location referred to , in order to "geo-tag" the uploaded document.

Many documents and metadata records refer to geographic locations in their text, yet do not tag these places with full geographical coordinates (e.g. a latitude/longitude reference) Geo-tagged metadata resources can be searched for using geographical queries, not just the "what", "when" and "who" queries of standard resource discovery interfaces. For example you could search a multimedia service asking: "Find all films that refer to locations within 50 km of Cologne", or search an on-line biological catalogue asking: "who is conducting research on species decline in Sardinia".

Part 3: Detailed, structured description.

Use Case Description	
Name	Geoparsing

Use Case Description	
Priority	medium
Description	Geoparsing refers to the process of identifying textual placename references within documents. The candidate placenames identified can then be verified against a gazetteer for purposes of either validation (spelling, variants) and/or for geocoding (i.e. attaching a coordinate reference to).
Pre-condition	A client application to submit resource (URI/external document) for processing. A Geoparsing Service which may avail of the uses of an integrated and/or external gazetteer, in which case a further precondition is the existence of: a suitable gazetteer lookup product/service. A client capable of portraying the geoparsed results – possibly as a map (which may utilise external WMS's for backdrop mapping) or as an XML document for further onward processing by client.
Flow of Events – Basic Path	
Step 1	User accesses client website and submits a URL/local document for processing and basic output method for results (map or XML).
Step 2	Geosparsing service is invoked by client and user submitted resource is automatically geoparsed.
Step 3 (optional)	Candidate placenames are referenced against a gazetteer (inline or service based) and the coordinate information extracted for each candidate place name (Note that a candidate may have multiple gazetteer entries and can thus be assigned multiple coordinates – an advanced geoparser may use this information to assist in disambiguating references, however in basic form all instances of candidates found in gazetteer are returned)
Step 4	Results delivered back to client for further processing
Step 5	User reviews output of geoparsed documents in client application and edits results achieved by automated process.
Step 6	Depending on options User either views results on a map or as an XML file. Former for visualisation purposes, latter for onward processing perhaps into a document management system.
Flow of Events – Alternative Paths	
Step 3a	No gazetteer referenced. Geoparsing results not georeferenced.
Step 6 a & b	Visualisation uses external WMS to provide background mapping to geoparser results. No visualisation, XML file of georeferenced place names is downloaded for offline processing e.g. to update spatial metadata related to original resource that was geoparsed.
Post-condition	A georeferenced interface file containing place names identified by the geoparser with appropriate georeferences (coordinate) information attached taken from a gazetteer.
User requirements derived from use case	
Content, Data, that is:	Clearly defined interfaces with charging policy and licensing regime clearly shown. Well documented data specification with maintained metadata.

Use Case Description	
▫ Relevant (minimal) completeness	Ideally large scale with variant and historical references. Global coverage would be ideal. Minimally, sufficient to deal with the resources being parsed – in European context that would be European coverage with full exonym/variant support.
▫ Name categories	Feature types are not a prerequisite but may be used by advanced parsing engines to assist in disambiguating features.
▫ Coverage	For EGN , sub optimally Europe
▫ Currency	Currency is dependent on the type of resources being parsed but most instances will require current (i.e. commonly within an annual update cycle) used terms. Certain applications may (ideally) however require historical references e.g. geoparsing historical reference collections such as the Statistical Accounts of Scotland [http://edina.ac.uk/statacc/]
▫ Spatial accuracy	Application dependent but no less than 1:50K tolerance
▫ Additional Info	Variant names, historical versions and multiple spatial footprints and feature types would be ideal. Hierarchical relationships not essential but could be useful for disambiguation takes as qualifiers.
Functionality	<in the form: input → output, for instance: place names → coordinate pairs (ideally full geometry) place names → feature type(s) place names → variant names
User Interface	n/a
Service Integration	<Requirements regarding EGN as Web service, e.g., protocol, usage of standards, performance, up-time> Protocol agnostic but ADL or OGC preferred End points available 24/7 Ability to real-time parse small document collections (hundreds of place terms). Offline facility for larger collections useful. Multiple presentation/outputs formats (geoparser) – maps, XML, CSV