CO-199

DEFINITION OF GUIDELINES FOR THE IMPLEMENTATION OF UNIQUE IDENTIFIERS AND LIFE-CYCLE INFORMATION IN PAN-EUROPEAN DATASETS

GREMEAUX N.

IGN France, SAINT-MANDÉ, FRANCE

ABSTRACT

The implementation of unique identifiers (UIDs) in pan-European datasets is one of the high priority demands of European users of the EuroGeographics products. It is also a growing concern for NMCAs since the publication of the INSPIRE directive in 2007 which requires the identification of geographical objects both in national and European datasets. ESDIN project’s WP9 partly focused on this issue. After an initial phase of investigation, a UID solution was proposed together with scenarios aiming at helping NMCAs implement this system in their national datasets while at the same time ensuring consistent deliveries of UIDs across Europe.

ACKNOWLEDGEMENTS

This paper is based on the work carried out as part of ESDIN WP9 and on the resulting deliverable D9.2-9.3 written by Nathalie DELATTRE (IGN Belgium), Gabriela DRAGAN (ANCPI Romania), Noémie GREMEAUX (IGN France), Kerstin I. JOHANSSON (NLS Sweden), Romain KERENEUR (IGN France), Johann THOMAS (IGN France), Esa TIAINEN (NLS Finland) and Luc Van Linden (1Spatial Great Britain).

1 BACKGROUND AND OBJECTIVES

1.1 Background

1.1.1 EuroGeographics and the pan-European products

EuroGeographics is a nonprofit organisation which gathers 52 National Mapping and Cadastral Agencies (NMCAs) from across 43 European Countries.

This organisation currently manages three pan-European datasets (EBM, ERM and EGM) which are made up of the national contributions of the member countries:

- EBM (EuroBoundaryMap) [1] is an administrative dataset at the scale of 1:100 000;
- ERM (EuroRegionalMap) [2] is a topographic reference dataset, which includes data about administrative units, transport networks, hydrography, settlements and named places, at the scale of 1:250 000;
- EGM (EuroGlobalMap) [3] includes the same data themes as ERM but at the scale of 1:1 000 000.

These datasets are updated and delivered every year to the European Commission. For each delivery, the national components are merged into a seamless database, with special care given to ensuring that the data is edge-matched on the international boundaries.

1.1.2 INSPIRE and ESDIN

In 2007, the European Commission published the INSPIRE directive, which aimed at harmonizing geographical data for environmental purposes across Europe. The Commission’s medium to long-term goal was for all databases in Europe, both national and pan-European, to conform to the INSPIRE requirements.

In this framework, INSPIRE-compliant specifications were written for different geographical themes, including some of the themes present in the EuroGeographics products – namely Administrative Units, Transport Network, Hydrography and Geographical Names.

In almost all feature classes, INSPIRE requires the implementation of an "external" identifier on each object, defined as a "unique object identifier, which is published by the responsible body, which may be used by external application to reference the spatial object". This identifier, called "inspireID", aims at uniquely identifying the geographical objects.

In order to facilitate the implementation of the directive among NMCAs, EuroGeographics launched the ESDIN project (European Spatial Data Infrastructure with a best practice Network) in 2008, so as to tackle a number of INSPIRE requirements and more particularly to define guidelines on how to integrate them into pan-European datasets.

The ESDIN project was divided into twelve work packages, each of which focused on a particular aspect of the maintenance of pan-European products in the INSPIRE framework. One of these work packages, WP9, dealt among other subjects with the issue of the implementation of unique identifiers in pan-European datasets.

As a result of this work package, the deliverable D92_93_ESDIN_Uids_guidelines_Final.pdf was published, in which guidelines were proposed to try and solve this issue. The present paper is based on the findings described in this deliverable.

1.2 Objectives

This paper intends to present the study and the results gathered in ESDIN WP9 regarding the implementation of unique identifiers in pan-European datasets.

The objective of this study was first to define what kind of unique identifiers were required at the pan-European level, given different user requirements, and to examine what was already done regarding the implementation of identifiers in European countries. The second objective was to propose an INSPIRE-compliant solution to implement and manage unique identifiers in pan-European products, regardless of the level of progress achieved by the NMCAs in their own national datasets.

2 APPROACH AND METHODS

“Unique identifier" is a rather broad concept, which can cover a number of specificities. The work in WP9 therefore started with a phase of investigation which aimed at defining the exact scope of the study and the orientation it should take.

2.1 Analysis of user requirements

An analysis of user requirements regarding pan-European datasets was first carried out. This analysis considered two kinds of users:

- pan-European users: the European Commission has notably shown in the past few years a growing interest in the implementation of unique identifiers in the EuroGeographics products;

- national users: this part of the analysis was more particularly based on IGN-F customers who had already been surveyed on the occasion of an internal project ("Echanges" project [4]).

The results of this analysis are summarized below. It appeared that the users especially needed to:

- visualize the modifications between two versions of a database;

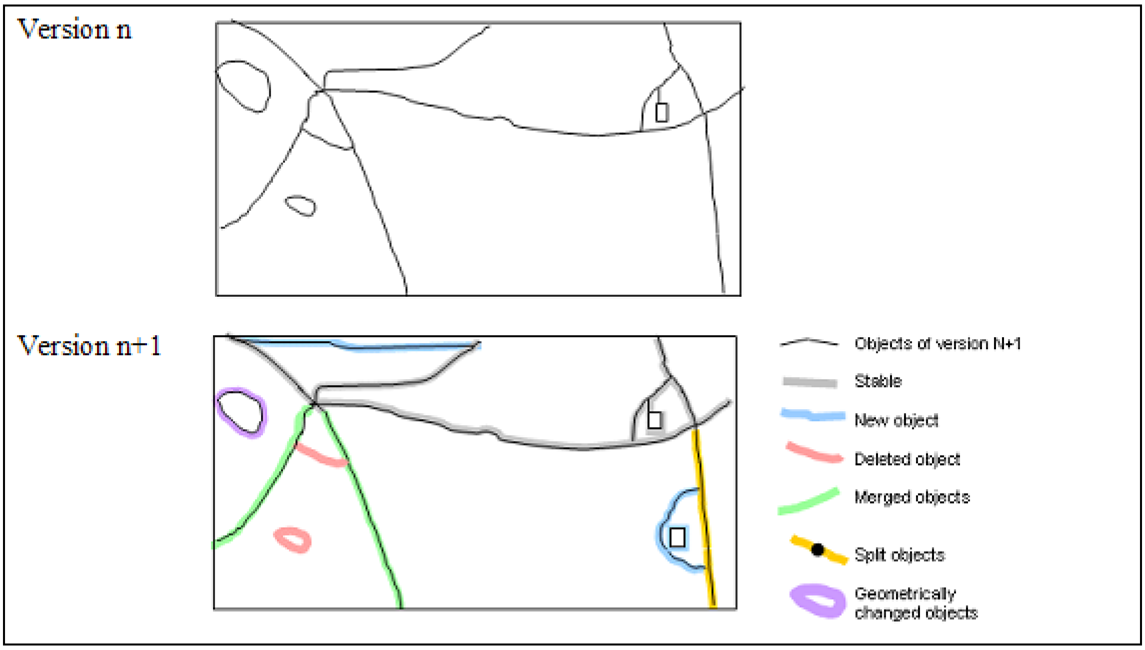


Figure 1: Visualization of modifications

- integrate data from multiple sources into a single database;

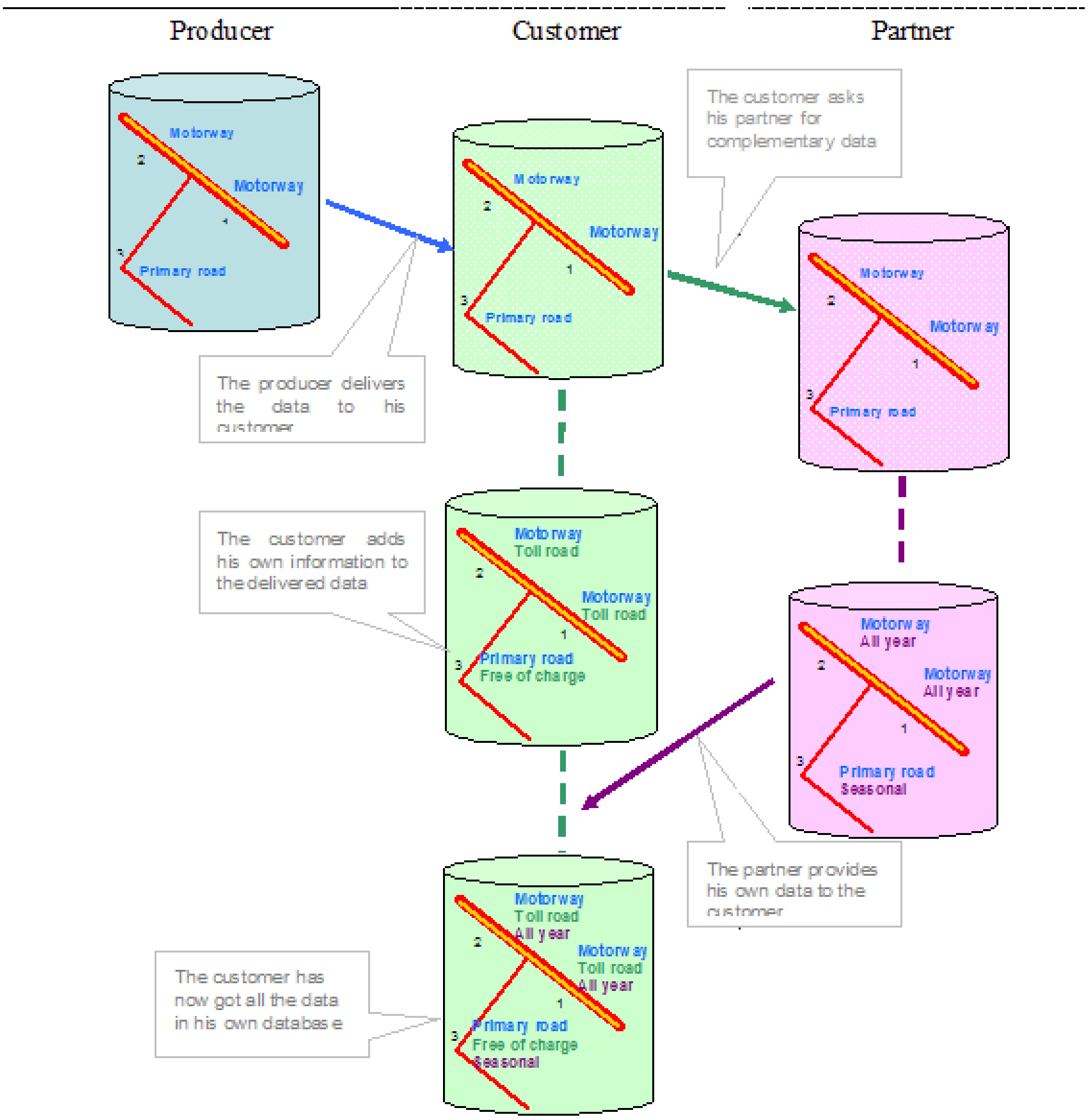
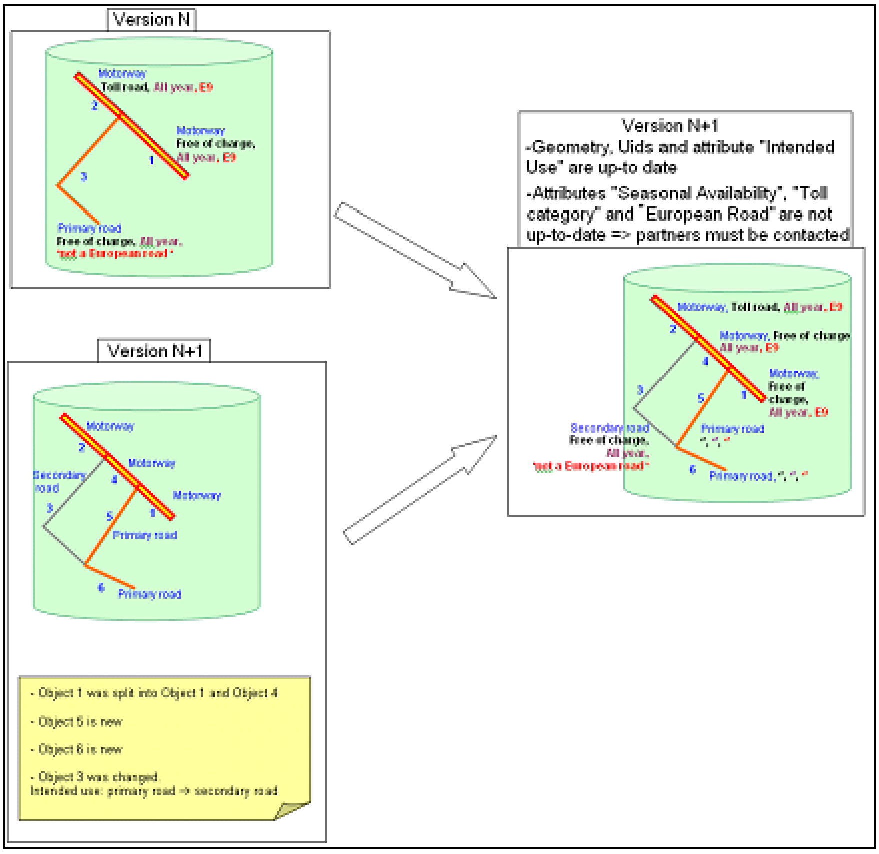

Producer
Customer
Partner
Motorway
Motorway
2
1
3
Primary road
The producer delivers the data to his customer
The customer asks his partner for complementary data
Motorway Toll road
Motorway Toll road
Primary road Free of charge
The customer adds his own information to the delivered data
Motorway All year
Motorway All year
Primary road Seasonal
The partner provides his own data to the customer
Motorway Toll road All year
Motorway Toll road All year
Primary road Free of charge Seasonal
The customer has now got all the data in his own database

Figure 2: Integration of data from multiple sources

- be provided “differential” information (i.e. information about the modifications between two releases) so as to update or transfer data to their own databases more easily.



-Geometry, Uids and attribute "Intended Use" are up-to date

-Attributes "Seasonal Availability", "Toll category" and "European Road" are not up-to-date => partners must be contacted

- Object 1 was split into Object 1 and Object 4

- Object 5 is new

- Object 6 is new

- Object 3 was changed.
Intended use: primary road => secondary road

Figure 3: Use of differential information during updates

It appeared from these results that the objects recorded in geographical datasets need to be clearly and reliably identified in order to answer the user requirements. The "external identifier" defined by INSPIRE would indeed correspond to their needs.

From now on, the term "UID" (Unique IDentifier) designates the "external identifier" according to the INSPIRE definition.

2.2 Overview of the possible modifications of geographical objects

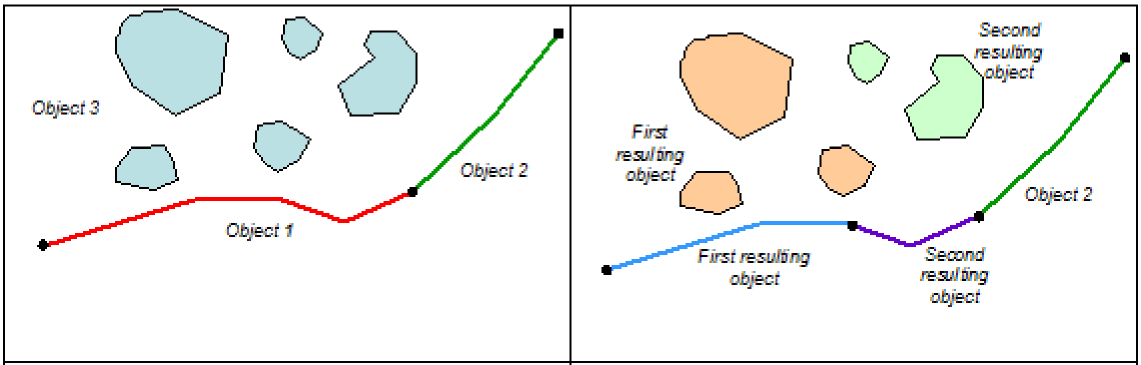
The second phase of the study was also based on the findings of the IGN-F project "Echanges" [4]. Ten types of possible modifications of geographical objects were identified:

- Creation: the object has been created, it did not exist in the previous version;

Suppression: the object existed in the previous version but has been deleted;

- Stability: the object has not been modified either geometrically or semantically since the previous version;

- Split: one object from the previous version has been split into two or more objects in the updated version;



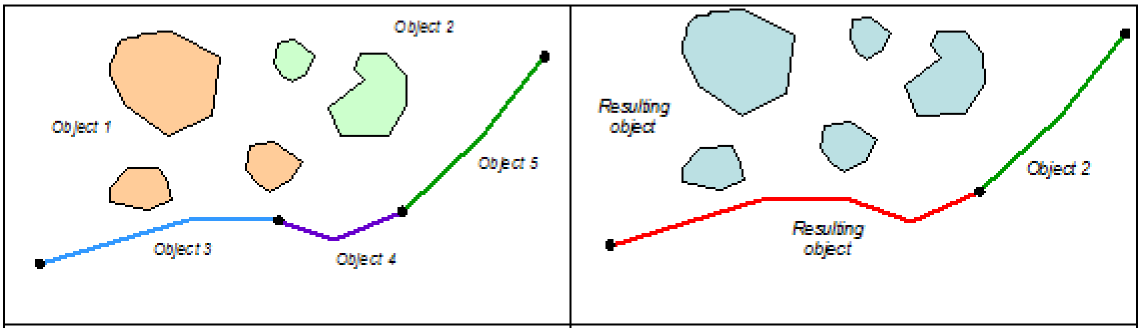
Version 1	Version 2

In this example, there have been two splits since the last version:

- the simple object 1 has been split into two simple objects;
- the multisurface object 3, originally composed of five surfaces, has been split into two multisurface objects.

Figure 4: Example of split features

- Merge: two or more objects from the previous version have been merged into a single object in the updated version;



Version 1	Version 2

In this example, there have been two merges since the last version:

- the simple objects 3 and 4 have been merged into a single simple object;
- the multisurface objects 1 and 2 have been merged into a single multisurface object, composed of five surfaces.

Figure 5: Example of merged features

- Geometric modification: the object's position and/or dimensions have been changed since the previous version. In the case of multisurfaces, the suppression or addition of some of the surfaces composing the multisurface is a geometric modification.
- Semantic modification: one or more attribute values of an object have been changed since the previous version;
- Mixed modification: the object has undergone both a geometric and a semantic modification since the previous version;
- Aggregate: objects from the previous version have undergone a series of splits and merges.
- Class modification: the object has been moved to another feature class.

The study led in WP9 tried to tackle both simple objects (points, lines and polygons) and complex ones (multipoints, multisurfaces…). However, it was only possible, due to lack of time, to go into detail for simple features, so this paper does not take complex objects into account. In what follows, only objects with simple geometries are considered.

2.3 Analysis of INSPIRE requirements

The INSPIRE specifications and Generic Conceptual Model [5] were then carefully studied in order to precisely understand the requirements regarding the "inspireID" mentioned in part 1.1.2. The main points of this analysis are briefly summarized here.

The inspireID needs to respect the following principles:

- Uniqueness: “No two spatial objects of spatial object types specified in INSPIRE application schemas may have the same external object identifier, i.e. the identifier has to be unique within all the spatial objects published in INSPIRE”. The uniqueness of external identifiers is ensured by the structure proposed by INSPIRE.

- Persistence: “The identifier has to remain unchanged during the life-time of a spatial object”.

- Traceability: “Since INSPIRE assumes a distributed, service-based SDI, a mechanism is required to find a spatial object based on its identifier. I.e. the identifier has to provide sufficient information about the source of the spatial object so that arrangements can be made that allow to determine the download service(s) that provide access to data from that source”.

- Feasibility: “The system has to be designed to allow that identifiers under existing national identifier systems can be mapped”.

INSPIRE also recommends that unique identifiers be composed of a namespace to identify the data source, followed by a local identifier, composed of letters and digits, unique within that namespace.

Three life-cycle attributes are associated to UIDs in the INSPIRE specifications:

- versionId: “the identifier of the particular version of the spatial object, with a maximum length of 25 characters. If the specification of a spatial object type with an external object identifier includes life-cycle information, the version identifier is used to distinguish between the different versions of a spatial object. Within the set of all versions of a spatial object, the version identifier is unique”.

- beginLifespanVersion: “date and time at which a version of an object was inserted or changed in the spatial data set”.

- endLifespanVersion: “date and time at which a version of an object was superseded or retired in the spatial dataset”.

2.4 Survey among NMCAs

Existing best practices regarding the implementation and maintenance of unique identifiers in Europe were also examined. A questionnaire was sent to different NMCAs across Europe, both in and out of the ESDIN project, in order to better estimate the level of achievement in different countries.

Five organizations provided answers to the questionnaire: Ordnance Survey Great Britain (OSGB), Kadaster Netherlands, BKG Germany, GEObasis Germany and IGN France.

The questions asked referred to the following themes:

- types of features identified by UIDs;
- UIDs structure;
- life-cycle rules (i.e. how UIDs are maintained when the objects are modified);
- existence and relevance of a versioning system to keep track of the modifications of the objects;
- impact for the users;
- impact on the data quality.

What has been gathered from this survey is that the INSPIRE principles regarding UIDs structure and characteristics were generally followed or could be adapted by the NMCAs who have implemented a unique identifier system. It was also obvious that different update rule systems had been adopted by the countries, and that a harmonized methodology would be needed at the European level to get homogeneous products.

2.5 Initial conclusion

As a result of this initial investigation phase, it was decided to follow the INSPIRE principles to devise the European UID solution: both the structure of the UIDs and the life-cycle information should comply with the INSPIRE requirements.

It also appeared that the solution needed to take into account the specificities of the pan-European products, which means specifying an adapted UID structure for transnational features and for features located on international boundaries.

Finally, the proposed solution had to take into account the very heterogeneous level of achievement of the NMCAs regarding UIDs across Europe.

3 RESULTS

3.1 UIDs and life-cycle information structure

The method presented above led to the definition of a structure for UIDs and life-cycle information in pan-European datasets.

3.1.1 UIDs structure

The structure was designed in order to ensure as much as possible the traceability of the identifiers. Each identifier, in accordance with the INSPIRE requirements, is composed of a namespace and a local identifier – made up of letters and digits – which is unique within that namespace. The namespace varies according to who manages the identifier and at what level (national or European). For countries which provide unique identifiers, the identifier provided for each object may in some cases be used as local identifier (see part 3.3 for the description of the cases).

As a result, a distinction is made between identifiers maintained by the countries and identifiers directly maintained at the European level, either because the countries are not able to provide them or because they identify objects created only when the pan-European dataset is assembled.

- Objects which already have a usable national UID

	Namespace	**Local identifier**
Objects with a national UID	• a prefix identifying the product or dataset (ExM, EBM, ERM, EGM); • the 2-character ICC code of the country the object belongs to; • an abbreviation indicating the data provider	The national identifier is used as local identifier

Example: an ExM object from IGN France could have the following identifier: ExM.FR.IGNF.BDCTRORO0000000019977628, where "BDCTRORO0000000019977628" is the identifier of this object in the French database.

- Objects whose local identifier has to be maintained at the European level

	Namespace	**Local identifier**
Objects located in one country only	• a prefix identifying the product or dataset (ExM, EBM, ERM, EGM); • the 2-character ICC code of the country the object belongs to; • the abbreviation "EGHO" to indicate that the ExM UID of the object is managed by EuroGeographics.	The local identifier is a 12-digit sequential number, generated at the European level.
Pan-European objects	• a prefix identifying the product or dataset (ExM, EBM, ERM, EGM); • the prefix "EU" to indicate that the object has been created to answer the specific needs of pan-European datasets and is directly managed at the European level; • the abbreviation "EGHO" to indicate that the ExM UID of the object is managed by EuroGeographics.	

Example: An ERM European Road (feature class ERoad) could have the following identifier: ERM.EU.EGHO.000000012548, where the local identifier 000000012548 is a sequential number directly created at the European level.

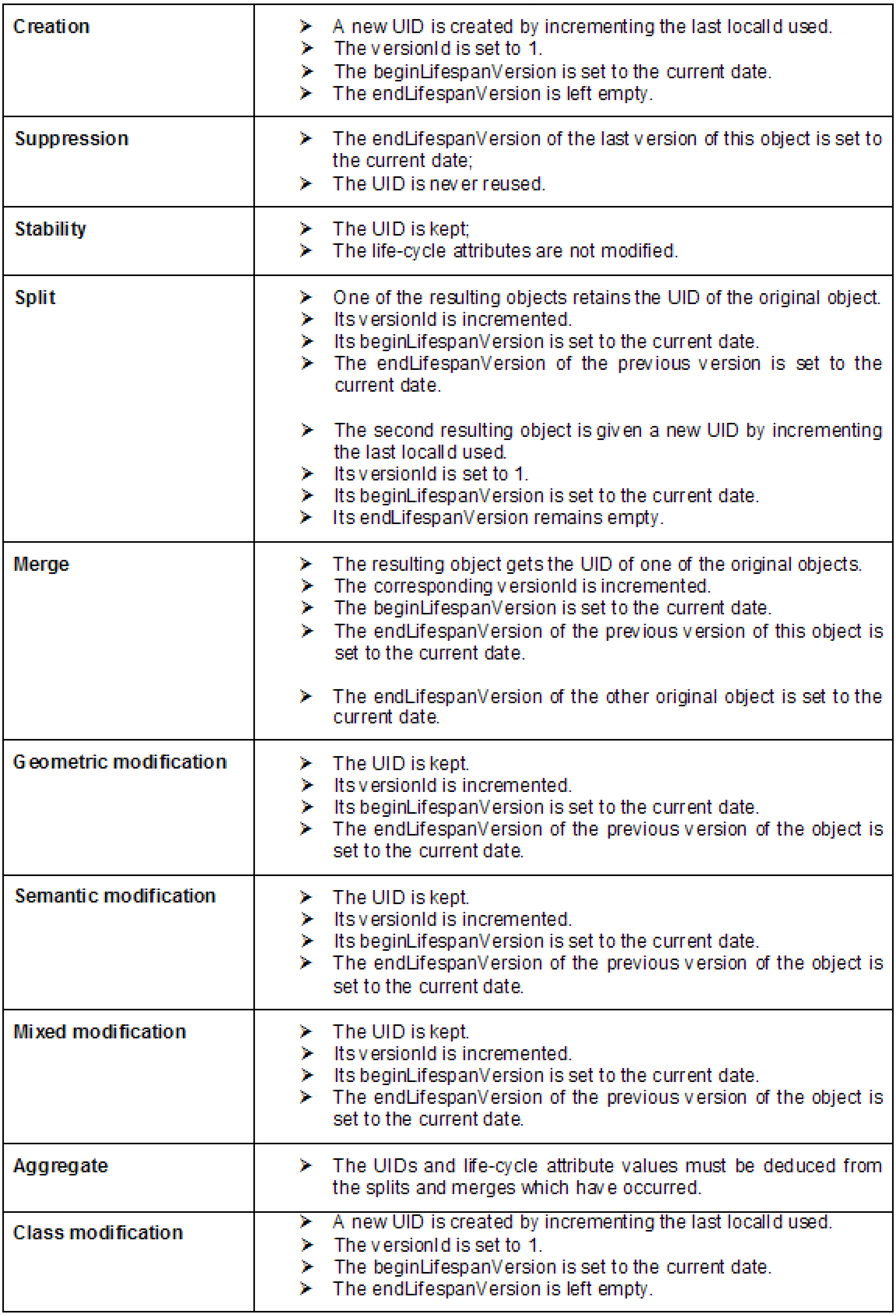
3.1.2 Life-cycle information

Three life-cycle attributes have been defined, both in conformance with INSPIRE and with the practices observed in the NMCAs who maintain this type of information:

Name	**Definition**
versionId	A number set to 1 when the object is added to the database, and incremented after each modification of the object
beginLifespanVersion	The date at which this version of the object was added in the database
endLifespanVersion	The date at which this version of the object was removed from the ExM database

3.2 ESDIN rules for managing UIDs and life-cycle information

Rules to manage UIDs and life-cycle information were also defined for each type of modification. These rules were also based on best practices observed in the surveyed NMCAs.



Creation	➤ A new UID is created by incrementing the last localId used. ➤ The versionId is set to 1. ➤ The beginLifespanVersion is set to the current date. ➤ The endLifespanVersion is left empty.
Suppression	➤ The endLifespanVersion of the last version of this object is set to the current date; ➤ The UID is never reused.
Stability	➤ The UID is kept; ➤ The life-cycle attributes are not modified.
Split	➤ One of the resulting objects retains the UID of the original object. ➤ Its versionId is incremented. ➤ Its beginLifespanVersion is set to the current date. ➤ The endLifespanVersion of the previous version is set to the current date. ➤ The second resulting object is given a new UID by incrementing the last localId used. ➤ Its versionId is set to 1. ➤ Its beginLifespanVersion is set to the current date. ➤ Its endLifespanVersion remains empty.
Merge	➤ The resulting object gets the UID of one of the original objects. ➤ The corresponding versionId is incremented. ➤ The beginLifespanVersion is set to the current date. ➤ The endLifespanVersion of the previous version of this object is set to the current date. ➤ The endLifespanVersion of the other original object is set to the current date.
Geometric modification	➤ The UID is kept. ➤ Its versionId is incremented. ➤ Its beginLifespanVersion is set to the current date. ➤ The endLifespanVersion of the previous version of the object is set to the current date.
Semantic modification	➤ The UID is kept. ➤ Its versionId is incremented. ➤ Its beginLifespanVersion is set to the current date. ➤ The endLifespanVersion of the previous version of the object is set to the current date.
Mixed modification	➤ The UID is kept. ➤ Its versionId is incremented. ➤ Its beginLifespanVersion is set to the current date. ➤ The endLifespanVersion of the previous version of the object is set to the current date.
Aggregate	➤ The UIDs and life-cycle attribute values must be deduced from the splits and merges which have occurred.
Class modification	➤ A new UID is created by incrementing the last localId used. ➤ The versionId is set to 1. ➤ The beginLifespanVersion is set to the current date. ➤ The endLifespanVersion is left empty.

3.3 Calculation methodology for maintaining UIDs and life-cycle information

The overall objective was to implement the ESDIN UID system and rules in the pan-European datasets regardless of what each NMCA could provide. As a result, four implementation scenarios have been defined:

Scenario 1: UIDs are not implemented at all at the national level

For the first delivery, UIDs are attributed to all the objects at the European level according to the ESDIN structure.

After each new delivery, the UIDs are calculated again at the European level by a comparison with the previous release.

1. For each object of the new delivery, the method consists in searching for an object from the previous release with the same attributes and geometry.

- If such an object is found, then the object has remained stable, and takes the former value of the UID.
- If an object with the same geometry but different attributes is found, then the object has undergone a semantic modification and the ESDIN rules are applied.
- Geometric modifications and mixed modifications are also detected this way.

2. Splits and merges are retrieved by a geometrical comparison of the objects not yet processed during step 1 between the two releases. The ESDIN rules are applied when necessary.

3. All the objects from the new delivery which have not been processed during steps 2 and 3 are considered as new objects and given a new UID. The objects from the former release whose UIDs have not been transferred are considered as deleted objects and the ESDIN rules are applied.

Scenario 2: UIDs are implemented at the national level but are not consistent with the ESDIN proposal

This scenario considers NMCAs:

- which have implemented UIDs but no life-cycle rules;
- which have implemented UIDs but cannot provide life-cycle information;
- whose UIDs do not follow the INSPIRE requirements of uniqueness, persistence and traceability.

There is no way to assess the quality of the UIDs provided. So as to ensure a perfect homogeneity and consistency across Europe, scenario 1 is applied.

A table recording the links between the national UIDs and the ones calculated for the pan-European products can be sent to the producer.

Scenario 3: UIDs are implemented at the national level and only partial life-cycle information can be provided

This scenario considers NMCAs where UIDs are implemented and follow the ESDIN rules, and life-cycle information can be provided for simple modifications (creation, suppression, and semantic, geometric or mixed modification).

The national UIDs of the objects for which life-cycle information has been provided are directly used as local identifiers in the European products. Steps 2 and 3 of scenario 1 are then performed.

Scenario 4: UIDs are implemented at the national level and complete life-cycle information can be provided

The national UIDs are directly used as local identifiers for the pan-European datasets.

The long-term objective is for all NMCAs to be able to follow scenario 4. In the meantime, the other scenarios ensure that UIDs can be consistently implemented in the pan-European products.

4 CONCLUSIONS AND PERSPECTIVES

The study presented here aimed at proposing a consistent approach towards the implementation of unique identifiers and life-cycle information in pan-European datasets, and more particularly in the three EuroGeographics pan-European products.

The work performed in WP9 started with an extensive investigation phase which allowed to better define the scope of the study. An analysis of both national and European users' requirements showed that what was currently most needed in the products was the implementation of external identifiers, also called UIDs in the document. In the context of the publication of the INSPIRE directive to harmonize data across Europe, it seemed only logical to try and apply the INSPIRE requirements while defining the UID solution. Existing UID systems and best practices of implementation among some NMCAs have also been studied through a survey.

This led to the proposal of a solution to implement INSPIRE-compliant UIDs and life-cycle information, together with update and maintenance rules to be applied homogeneously across Europe. The proposed system can be managed both at the national and at the European level. Four scenarios have been developed to enable all NMCAs to progressively reach the same level of achievement regarding UIDs in their national datasets, while at the same time ensuring consistent delivery of UIDs in the pan-European products.

However, this whole work remains for now at a theoretical level. It is expected that the proposed solution might evolve as some unpredicted issues are bound to appear during the implementation phase.

The production of the three EuroGeographics pan-European products – EBM, ERM and EGM – will go on at least for the next four years thanks to the renewal of the contract with the European Commission. The Commission's request for UIDs is now a high priority. NMCAs will therefore very soon have the opportunity to test the proposals presented in this paper.

APPENDIX A – REFERENCES

[1] EuroBoundaryMap Project Overview, EuroGeographics website, 2009,
http://www.eurogeographics.org/content/euroboundarymap.

[2] EuroRegionalMap Project Overview, EuroGeographics website, 2009,
http://www.eurogeographics.org/content/euroregionalmap-0.

[3] EuroGlobalMap Project Overview, EuroGeographics website, 2009, http://www.eurogeographics.org/content/euroglobalmap.

[4] IGN-France project “Echanges” website, 2009,
http://projet-echanges.ign.fr/

[5] INSPIRE (2009): INSPIRE Generic Conceptual Model, Data Specifications Drafting Team, version 2009-08-26
http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2