Extending the Intermediate Data Structure (IDS) for longitudinal historical databases to include geographic data

Hedefalk, Finn; Harrie, Lars; Svensson, Patrick

Published in:
Historical Life Course Studies

2014

HISTORICAL LIFE COURSE STUDIES

VOLUME 1
2014

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at http://www.ehps-net.eu/journal.

Editors: Koen Matthijs & Paul Puschmann
Family and Population Studies
KU Leuven, Belgium
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 1 (2014), 27-46, published 05-09-2014

Extending the Intermediate Data Structure (IDS) for longitudinal historical databases to include geographic data

Finn Hedefalk
Department of Physical Geography and Ecosystem Science, Lund University

Lars Harrie
Department of Physical Geography and Ecosystem Science, Lund University

Patrick Svensson
Centre for Economic Demography, Lund University

ABSTRACT

The Intermediate Data Structure (IDS) is a standardised database structure for longitudinal historical databases. Such a common structure facilitates data sharing and comparative research. In this study, we propose an extended version of IDS, named IDS-Geo, that also includes geographic data. The geographic data that will be stored in IDS-Geo are primarily buildings and/or property units, and the purpose of these geographic data is mainly to link individuals to places in space. When we want to assign such detailed spatial locations to individuals (in times before there were any detailed house addresses available), we often have to create tailored geographic datasets. In those cases, there are benefits of storing geographic data in the same structure as the demographic data. Moreover, we propose the export of data from IDS-Geo using an eXtensible Markup Language (XML) Schema. IDS-Geo is implemented in a case study using historical property units, for the period 1804 to 1913, stored in a geographically extended version of the Scanian Economic Demographic Database (SEDD). To fit into the IDS-Geo data structure, we included an object lifeline representation of all of the property units (based on the snapshot time representation of single historical maps and poll-tax registers). The case study verifies that the IDS-Geo model is capable of handling geographic data that can be linked to demographic data.

e-ISSN: 2352-6343
PID article: http://hdl.handle.net/10622/23526343-2014-0003?locatt=view:master

The article can be downloaded from here.

1 INTRODUCTION

Several projects and countries are creating demographic databases with an Intermediate Data Structure (IDS) as part of the European Historical Population Samples Network (EHPS-Net) (Alter & Mandemakers 2014). In total, EHPS-Net aims to convert, at minimum, 15 longitudinal historical databases into a common database schema. This IDS schema includes individuals and contexts (e.g., geographic places), as well as the relationships between them. Data extraction programs will thereafter be developed to transform the data into usable datasets for longitudinal analysis. Furthermore, metadata and an ontology are created to enable the interpretation of data from different countries and communities (EHPS-Net 2009). Researchers as well as the public will then have access to data in a standardised structure explained by metadata. This will facilitate international data sharing and comparative research.

To incorporate geographic factors in the demographic research, it is vital that all individuals are linked to a physical location. For newer demographic data, detailed address information can in many cases be linked directly to GIS-databases. However, for older demographic data there are seldom addresses available and linkage has to rely on identifying objects and assigning them the proper geographical coordinates. These geographic data are primarily property units and buildings. This means that we often have to create tailored geographic datasets. In those cases, there are benefits of storing the geographic data in the same structure as the demographic data. In particular, the usability of the data should increase when they are exported and distributed as integrated datasets. The links between the demographic and geographic data may also be easier to maintain. Finally, since the main aim of IDS is to function as a standardized distribution format for historical longitudinal data, geographic data, that are integrated with such data, need to be described in a standardized way as well. Hence, there is a need of a geographically extended version of IDS.

When individuals are linked to such locations, it is possible to investigate the environment in which they lived. Having access to such integrated micro-level geographic and historical demographic data allows us to more deeply and accurately understand the geographic factors that affected people throughout their history. When having the spatial locations of each individual, we can, for instance, analyse the impacts that land reforms had on the health of different social groups through changes in population density, and thereby exposure to diseases. By adding external geographic data, we may also analyse how soil and elevation conditions affected production at the farm level, and how proximity to gathering points influenced economic development.

IDS version 4 model (Alter & Mandemakers 2014) includes the possibility to link geographic data to IDS and store point data within IDS. The aim of this study is to create the possibility to integrate the geographic data within IDS in a new model, coined IDS-Geo. The first part of the paper describes our proposed changes in the IDS data structure, as well as, geometric data types. The second part of the article includes a case study in which we implement IDS-Geo and populate it with data from the SEDD IDS-Geo v.1 database. This database is a geographically extended version of the Scanian Economic Demographic Database (SEDD) created by the Centre for Economic Demography (CED) at Lund University (Bengtsson, Dribe & Svensson 2012). With this study we do not intend to create a new geographic branch of IDS. Instead, we hope that it could influence future development of the IDS structure.

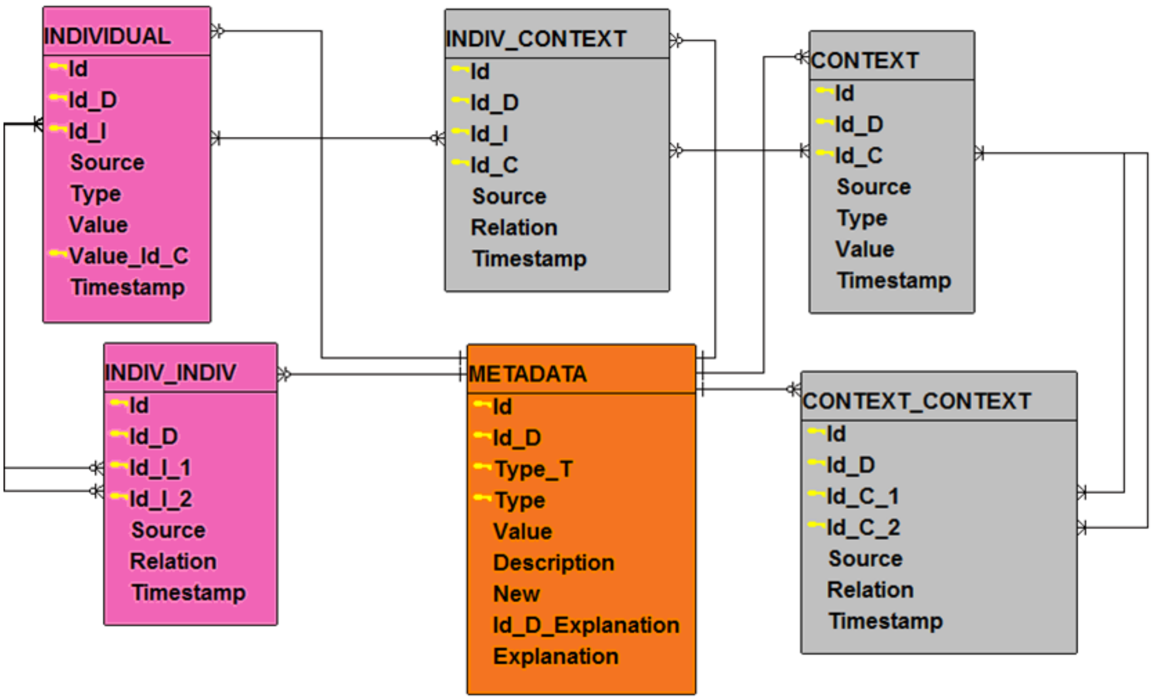
A note on the terminology used in this paper: we use the word entity to describe things that exist in the world and that can be distinguished from each other. Furthermore, an entity may be either an object or an event. We define objects as concrete and lasting entities, such as buildings or people; and events as things that happen, are instantaneous or exist for a period of time, and then disappear (e.g., births, constructions of buildings, or enclosures rearranging property units) (Grenon & Smith 2004).

2 IDS FOR LONGITUDINAL HISTORICAL MICRODATA

The expected outcome of EHPS-Net is that the administrators of historical longitudinal databases will transform their data into the common data structure IDS. EHPS-net is also creating a portal that gathers metadata about the databases, as well as provides data extraction programs (EHPS-Net 2009). The following tables are used in IDS version 4 (Alter & Mandemakers 2014):

- INDIVIDUAL – Stores observations about individuals. Each observation forms a row.
- CONTEXT – Stores observations about the context where the person lived. This entails human information (such as households), administrative information (parishes, property units, etc.) as well as information about the physical environment (buildings, etc.).
- INDIV_INDIV – Contains relationships between two persons (e.g., marriages, hierarchical relations or family relations).
- INDIV_CONTEXT – Links a person to a context, either momentarily or during a time period (e.g., a person living in a household).
- CONTEXT_CONTEXT – Contains relationships among objects in the CONTEXT table, for example, a household within a municipality.
- METADATA – Contains code lists that define all the variables and values used, which explains, for example, the codes used in the Type and Relation attributes.

Figure 1 *Tables used in IDS version 4*



Source: Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4. Historical Life Course Studies, 1, p.17.

Figure 1 illustrates the tables and their attributes. The INDIVIDUAL and CONTEXT tables contain a set of attributes for the individuals and the contexts in which they lived. Each row in the tables only contains one attribute value (in the Value attribute), where the Type attribute value explains what type of information is stored. Furthermore, the Timestamp attribute specifies an exact time or a time period for the attribute value and the Source attribute contain data about the source of the information. The association tables (INDIV_INDIV, INDIV_CONTEXT and CONTEXT_CONTEXT) have a similar structure; in this case, the Relation attribute specifies what type of relation the objects (e.g., Id_I_1 and Id_I_2 in the INDIV_INDIV table) have. Finally, the code Id_Polygon in the Type attribute (in CONTEXT) makes it possible to establish links to external geographic data.

The INDIVIDUAL table is able to store longitudinal data on an individual level. That is, it can contain life histories where a process in a person's life is stored with a start time and an end time. This general structure allows for storing both objects and events, which is essential for historical studies (Alter, Mandemakers & Gutmann 2009).

3 IDS-GEO: A GEOGRAPHICALLY EXTENDED IDS DATA MODEL

This section contains a description of the geographically extended IDS, named IDS-Geo. We have chosen to describe IDS-Geo using a conceptual data model (see e.g., Yeung & Hall 2007) that is independent of the type of database management system. The conceptual model is designed in Unified Modeling Language (UML). We also specify an eXtensible Markup Language (XML) Schema for IDS-Geo, which is a low-level logical model for data distribution. In this schema, the geographic elements are specified by the XML grammar Geography Markup Language (GML). One important purpose of IDS-Geo is the possibility of extracting IDS data easily into the IDS-Geo model, mainly the XML Schema. Therefore it is important to make it possible to establish one-to-one mappings between the models (e.g. that the CONTEXT attribute Type in IDS have a single correspondence in the IDS-Geo model). However, before introducing the IDS-Geo model, we have to describe the structure and time representation of the geographic data.

3.1 GEOGRAPHIC DATA

The first point to consider is what types of geographic data are required. The aim of the IDS structure is primarily to facilitate the exchange of historical demographic data; that is, we only require geographic data to link individuals to spatial locations. If address data are not available, this linkage can be done using, for example, buildings and property units. That is, IDS should support storage of, or linkage to, geometric representations of these objects. Other types of geographic data (e.g., soil type and roads) could be used in a longitudinal analysis without being stored within the IDS, as long as standard spatial reference systems are used.

The second point to consider is whether we should link to external geographic data or integrate the data within IDS. The CONTEXT table in IDS version 3 and 4 uses the Type code Id_Polygon, which enables linkage to external geographic data (Alter & Mandemakers 2012; 2014). An advantage with this approach is that you do not need to store geographic data within IDS. This is of special interest for geographic data that have standardized identifications; examples of such identifications are modern addresses and building identification numbers. In these cases we can link the IDS data to several geographic databases. However, for older geographic data such standardized identifications are often missing and there are seldom good external geographic databases. In particular, suitable databases with object lifeline or event chronicle time representations of historical geographic data are rare. In these cases we often have to create a tailored geographic dataset for the demographic data (cf. section 4). We argue that in those cases where there are tailored geographic data, they should be stored within the IDS. IDS version 4 supports storage of geographic point data, but in order to properly store building and property unit data, we also need to store line and polygon data; that is, we have to add more geometric data types.

The third point to consider is which geometric representation should be used. We propose that IDS-Geo should use the geometry data types in the Simple Feature Geometry object model defined in Simple Feature Access - Part 1: Common Architecture (Herring 2011). This standard was originally proposed by the Open Geospatial Consortium, but it has also been approved by the International Standard Organization (ISO 19125-1). Simple Feature Access describes, in a platform-independent way, a common architecture for handling spatial data. The core part is the geometry object model, which is a UML model of the spatial entities. The specifications are limited to 0, 1, and 2-dimensional vector geometry, and the main geometry types are points, line strings and polygons. This means that the standard does not include 3D geometries, coverages (e.g., raster data) or complex curves (e.g., Bezier curves). Simple Feature Access is implemented by most of the database management systems (DBMs) supporting geographic objects. Moreover, each geometry object in the Simple Feature Access standard is linked to a spatial reference system through use of Spatial Reference System Identifiers (SRIDs).

3.2 TIME REPRESENTATION OF GEOGRAPHIC DATA

We also need to consider the time representation of the geographic data. In this section we distinguish three types of basic time representations: temporal snapshots, object lifelines and event chronicles (Worboys 2005). Below follow a description of these three models as well as examples. The examples are given in Tables 1-3 and are loosely based on SEDD IDS-Geo v.1 (see section 4.1).

Sources for historical geographic data are often scanned historical maps, which can be regarded as snapshots of the conditions at a certain time. From such historical maps, objects such as property units and buildings can be digitised. Thus, one of the simplest models for storing spatio-temporal data is to assign each digitised object a timestamp corresponding to the date of the historical map (Table 1). Models for storing such time stamped objects are usually called snapshot models (Armstrong 1988) or temporal snapshots (Worboys 2005). Temporal snapshots are simple to create, but they are not suited for tracing objects through time and for detecting change. In Table 1, a property unit named "Hög 5" has been digitised (from three historical maps) and stored as three different objects. Each object is assigned a timestamp that represents the creation date of the historical map. Note that the geometry has changed in the two latter rows (indicated by polygon 2a and polygon 2b). The reason for this change is that an area of the property unit has been subdivided. Also note that the geometry is not exactly the same in the two latter rows (indicated by indexes a and b). This is because the rows are based on different maps with non-perfect geometries.

Table 1 *Temporal snapshots of the property unit 'Hög 5' (spelled 'Höj 5' in some sources).*

Id	name	timeStamp	geometry
15	Höj 5	1804-01-01	*(polygon 1)*
22	Hög 5	1820-01-01	*(polygon 2a)*
89	Hög 5	1865-01-01	*(polygon 2b)*

Explanation: The timeStamp attribute represents the creation date of each historical map.

Table 1 does not store the specific time of change for the objects. To enable such tracing object lifelines (Table 2) can be used (Worboys & Duckham 2004). In this time representation model, each object is assigned a time period for when it is valid in the real world. For example, a property unit exists from its construction to its destruction. In Table 2, the objects in Table 1 have been linked to a common identifier. Moreover, additional textual sources have been used to attain a more precise estimation of the time period during which the property unit and its geometries existed in the real world.

Table 2 *The property unit 'Hög 5' stored as object lifelines.*

id	propertyUnitId	name	startDate	endDate	geometry
4	pu_hog_52	Hög 5	1790-08-01	1815-08-01	*(polygon 1)*
5	pu_hog_52	Hög 5	1815-08-01	1890-08-01	*(polygon 2b)*

Explanation: startDate and endDate represent the valid time period of the object.

In Table 2, the geometry polygon 2b has been used instead of polygon 2a. The reason is that the source data for this polygon are of better quality than the ones for polygon 2a. Object lifelines are widely used in the GIS domain, but to better address the behaviours and interactive relationships of spatio-temporal objects and events, the event chronicles model (Table 3) has been proposed (Worboys 2005). Here, the focus shifts from the objects to the events. That is, instead of describing the states of objects, the events affecting the objects, as well as each other, are described. Table 3 illustrates how events that affected the property unit Hög 5 can be represented. In this example, the property unit was created on 1790-08-01. Then, a part of Hög 5 was subdivided on 1815-08-01 into the new property unit Hög 5a. Finally, it was partitioned on 1890-08-01 into two new property units: Hög 5b and Hög 5c (see Figure 2 for further explanation of the terms subdivision and partition). If such information about a property unit is available, storing it as an event chronicle could allow for a more detailed description of events and objects. Note that Table 3 shows a simplified example. To better model a large process (in which many events are involved) as a whole, we could describe the events in more detail, as well as model the relationships between the events themselves and how they are involved in the process (Yuan & Hornsby 2008).

Table 3 *Events linked to the property unit Hög 5 stored as event chronicles.*

eventId	eventName	date	propertyUnit
43	Created	1790-08-01	Hög 5
25	Subdivided	1815-08-01	Hög 5; Hög 5a
69	Partitioned	1890-08-01	Hög 5 -> Hög 5b; Hög 5bc

Explanation: For readability, we use names in the propertyUnit attribute instead of identifiers.

Both the temporal snapshots and object lifelines representations are widely used in historical GIS databases (Vanhaute 2003; Berman 2003; Dam 2013; Fitch & Ruggles 2003; Gregory & Southall 2005). The latter model is also common in standardisation work for geographic data. For example, the 34 data specifications created within the INSPIRE Directive use object lifelines (INSPIRE 2013). For longitudinal historical databases, the focus is on individuals, as well as on the events affecting them (Alter et al. 2009). Thus, these databases more often use combinations of object lifelines and event chronicles. This is also true for IDS version 4. By using a very generic structure, it allows for storage of temporal snapshots, object lifelines and event chronicles (Alter & Mandemakers 2014). Because IDS-Geo should enable storage of both historical geographic data and demographic data, it needs to allow for a combination of object lifelines and event chronicles.

3.3 CONSIDERATIONS OF STORING OBJECT LIFELINES IN IDS-GEO

In this section we describe three issues when storing object lifelines using the IDS-Geo schema: (1) uncertainty intervals in the lifelines; (2) object change; and (3) representing lineage relationships. It should be noted that the main part of the discussion is also applicable for IDS.

Common sources for historical geographic data are snapshot data (historical maps), and therefore several uncertainties in the object's lifelines representation can occur. This is illustrated in the following example. Two geometries (A and B) for a property unit have been digitised from two historical maps from 1800 and 1890. When storing the time of the geometries in its simplest form, the creation dates of the maps are used as dates for the geometries (Table 4).

Table 4 *Digitised geometries of a property unit*

Id	propertyUnitId	geometry	date
1	pu_hog_60	A	1800
2	pu_hog_60	B	1890

When we want to link individuals to their property units, a more detailed estimation of the objects' lifelines is needed. One way to do this is to link any existing data on properties, for example annual poll-tax registers, to the geometries. Then we can use, for example, changes in property units in these registers to better estimate the objects' lifelines. In our example, we know that geometry A exists from 1790 to at least 1820 and that geometry B exists from, at latest, 1840 to 1910. However, there is an uncertainty for the period 1820–1840 caused by, for example, incomplete or uncertain register information; geometry A may exist as late as 1840, whereas geometry B may exist as early as 1820. That is, the geometries have one maximum interval and one minimum interval. Such uncertainty intervals need to be represented in the database. Table 5 shows an example of storing the uncertainty intervals of both start dates and end dates.

Table 5 *Start and end date intervals for the two geometries of the property unit in Table 4*

Id	propertyUnitId	geometry	startDateMin	startDate	endDate	endDateMax
1	po_hog_60	A	1790	1790	1820	1840
2	po_hog_60	B	1820	1840	1910	1910

In IDS version 4, the Timestamp attribute provides an interval of a date (i.e., Start_date and End_date) but not intervals of the Start_date and End_date themselves. Nevertheless, there are alternative ways of storing such intervals. Table 6 shows an example of how to present the information from Table 5 in IDS-Geo, using the Timestamp attribute from IDS version 4. In Table 6, there are four observations of two geometries (A and B) for context po_hog_60. The Value attribute contains links to the two geometries stored in an external table. The start date and end date intervals are represented by the values intervalMinimum and intervalMaximum in the IDS Estimation attribute (these could be added to the IDS version 4 metadata table). That is, intervalMinimum represents the shortest date interval, whereas intervalMaximum represents the longest date interval.

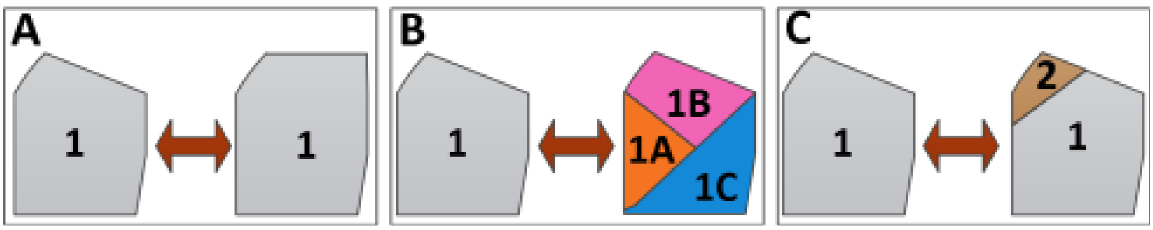
Table 6 *A CONTEXT table storing uncertainty intervals using the Timestamp attribute Estimation*

Id	Id_C	Type	Value	Start_date	End_date	Estimation
1	po_hog_60	geometry	A	1790	1820	intervalMinimum
2	po_hog_60	geometry	A	1790	1840	intervalMaximum
3	po_hog_60	geometry	B	1840	1910	intervalMinimum
4	po_hog_60	geometry	B	1820	1910	intervalMaximum

Explanation: The Estimation attribute could also be used to describe data providers' estimations of which dates they believe are the most probable ones.

Another issue to handle is the changes that occur to an object during its lifeline. Except for creation and destruction of an object, common types of changes for historical geographic data are geometry change, merge or split, and extraction or absorption (Worboys & Duckham 2004). In land surveying, for example, we have the following types of changes to property units: reallotment (Figure 2a), partitioning (Figure 2b) and subdivision (Figure 2c). An important issue is for which changes the identification of the property unit remains. In Figure 2 we have used the rules for identity in the Swedish cadastre system; other countries have other rules (e.g. that a partitioning leads to two new property units).

Figure 2 *Common types of changes that can occur to property units in historical data*



Explanation: The numbers inside the polygons represent object identifiers. (A) Here, a geometric change has occurred. The reason could be a reallotment; for example, new land is added to the property unit without affecting adjacent property units or creating new ones. Another possible reason is that the polygons have been digitised from historical maps with non-perfect geometry (i.e., the geometric change did not happen in the real world); (B) one property unit is partitioned into three new units, or three property units are merged into one new unit; (C) a part of property unit 1 is subdivided into the new property unit 2, or property unit 1 absorbs property unit 2. A principal difference between B and C is that in the former the property units lose their identities, whereas in the latter one, the property unit keeps its identity.

A geometric change that does not create or destroy any objects is simple to represent using the IDS schema. For example, the change in Figure 2a can be stored as two geometric observations of the object. However, when objects are being created or destroyed (Figure 2b-c), it is important to build links between the successors and predecessors. For example, if object 1 is split into the objects 1A, 1B and 1C, we may want to be able to trace the history and relationships between them. This can be necessary when linking demographic data to geographic data, or when analysing the evolution of objects through time. When storing such relationships, we propose using the CONTEXT_CONTEXT table and the term predecessor (and/or successor) for the Relation attribute (Table 7).

Table 7 *Using the CONTEXT_CONTEXT table to store the relationships between a predecessor (Id_C_1: 1) and its successors (Id_C_2: 1A, 1B, 1C).*

Id	Id_C_1	Id_C_2	Relation	Date
1	1	1A	predecessor	1840
2	1	1B	predecessor	1840
3	1	1C	predecessor	1840

Finally, it is important that the information stored in the IDS CONTEXT table is using the timestamp in a similar way of representing the time dimension as in the INDIVIDUAL table. This could, for example, be that a geographic record of a building is stored with a start date (the earliest documentation of the building) and an end date (the last documentation of the building). If information in both the INDIVIDUAL table and the CONTEXT table are representing time in such a way, it will be possible to model processes in time in the IDS INDIV_CONTEXT table. For example, a person lived in a building from a certain start date to a certain end date.

3.4 IDS-GEO CONCEPTUAL UML MODEL

Figure 3 shows a conceptual UML model of IDS-Geo. There are three main differences from the IDS version 4 model. The first difference is that IDS-Geo has separate classes for entities and the observations of the entities. This implies that the classes INDIVIDUAL and CONTEXT from IDS version 4 are split into two classes each: INDIVIDUAL and INDIVIDUAL_ENTITY, and CONTEXT and CONTEXT_ENTITY. INDIVIDUAL_ENTITY and CONTEXT_ENTITY have only one row for each entity, which includes a unique identifier. The classes INDIVIDUAL and CONTEXT represent observations of entities and contain all of the attributes that are included in the IDS version 4 classes INDIVIDUAL and CONTEXT tables. The associations between the features are realised as association classes (i.e., CONTEXT_CONTEXT, INDIV_CONTEXT and INDIV_INDIV). Lastly, all the Timestamp attributes use the data type Timestamp.

The reasons for separating the entities from their observations are as follows:

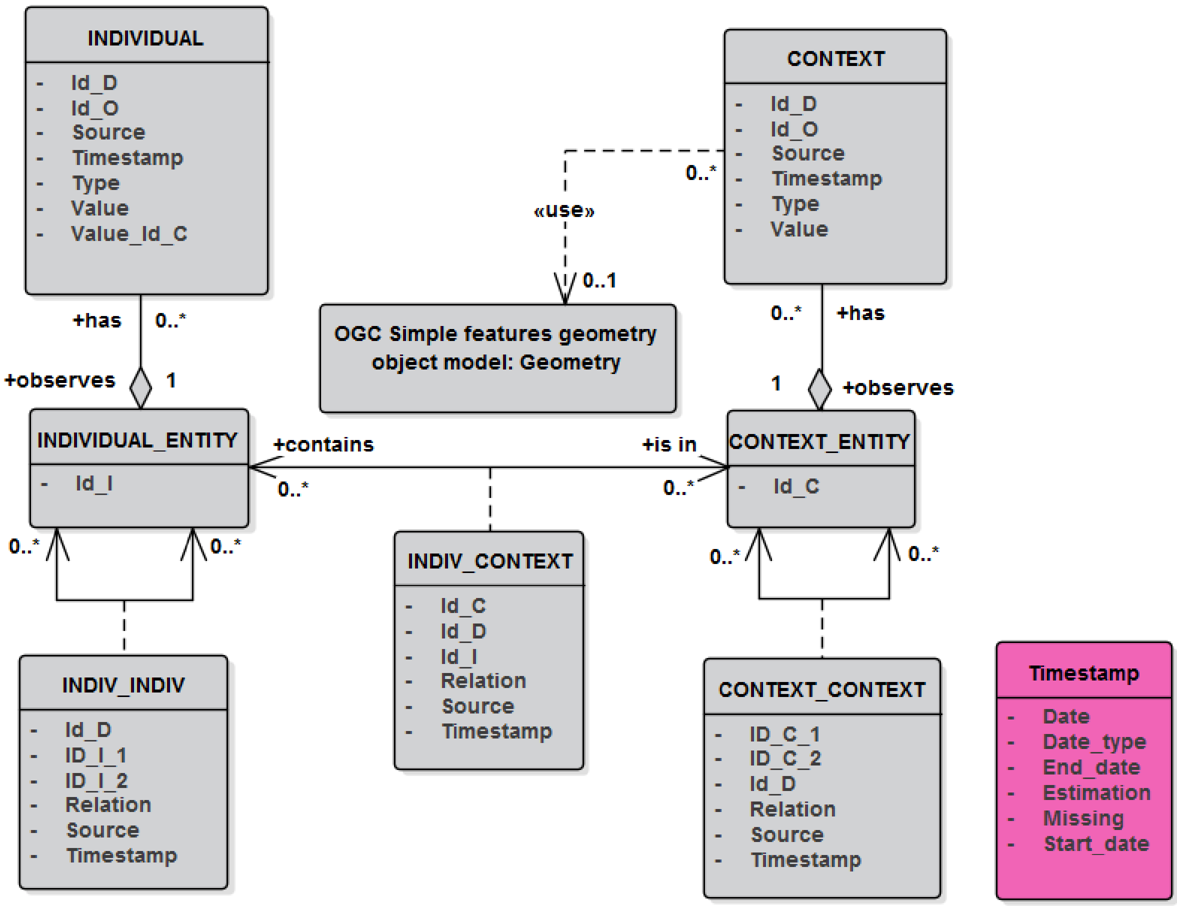
- It is more logical because they are separate classes in our conceptualisation of the world (i.e., in the real world there exist entities that may be observed).

- Some queries may be easier to construct due to the split, and it may be easier to create and represent the entities' relationships with each other.

- Coherence with the ISO 19156 standard on Observations and Measurements (O&M). O&M provides a standard schema and format for representing and exchanging various types of observations, as well as for the entities involved in them. This standard is important for data discovery and quality estimation (Cox, 2010). The latter allows the users to determine the usability of the data. A similarity between the IDS and O&M can be found in the following description. Walker et al. (2009) define an observation as "… an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure" (p. 4384). The result may be a number, term or other symbol, and a procedure may, for example, be a sensor, instrument, observer or algorithm (Cox, 2010). This view of an observation is very similar to the one used in IDS. That is, an observation of a specific context or individual can be seen as an action whose value is an estimate of the value of some type of context or individual, obtained from a specified source. Due to the many similarities, it is important to consider O&M and similar standards when creating the IDS-Geo structure. One reason for this is that because O&M is a published standard, it has gone through a rigorous testing and evaluation process (ISO/IEC, 2014). Another is that IDS-Geo data may in the future be combined with observation data from other domains. Note also that the O&M schema differentiates between the observed entity and the observation (that is, it uses individual classes for them). On the other hand, due to the general structure of IDS, it should be simple to convert IDS/IDS-Geo data to such standards without separating between the entity and the observation.

One disadvantage, however, is that the split creates more classes, which adds complexity to the model.

Figure 3 *Conceptual UML model of IDS-Geo.*



Explanation: The following multiplicity rules are applied. One entity class (INDIVIDUAL_ENTITY or CONTEXT_ENTITY) may have zero or more observations, whereas one observation may observe only one entity at a time. The relationships among the entities may be zero or more. Lastly, the geometry of an observed entity may be zero (i.e., without a geometry) or one (i.e., each observation may have only one geometry). A geometry object, however, may be observed multiple times. In this conceptual model, no primary and foreign key constraints are modelled. These are added later on in the logical models.

The differentiation of observations and entities in the conceptual model should not complicate the conversions between implementations of IDS and IDS-Geo. For example, in a physical database implementation, it is still possible to link the observation tables with each other without using the entity tables which can be omitted in the join queries (that is, the observation tables can be equivalent to the CONTEXT and INDIVIDUAL tables in IDS). Users can therefore query a relational database implementation according to IDS version 4. This should also hold true for the specified IDS-Geo XML Schema (section 3.5), because it is possible to establish a one-to-one mapping when exporting IDS data to the XML Schema.

A second difference of IDS-Geo from IDS version 4 is the absence of a metadata class. The reason is to avoid possible confusion of having two metadata tables, and it is therefore better to use the IDS metadata table for IDS-Geo data. However, instead of using downloadable metadata tables, a central online code register may facilitate the update and maintenance of the metadata (see e.g. EC 2014).

3.5 IDS-GEO DATA XML EXPORT

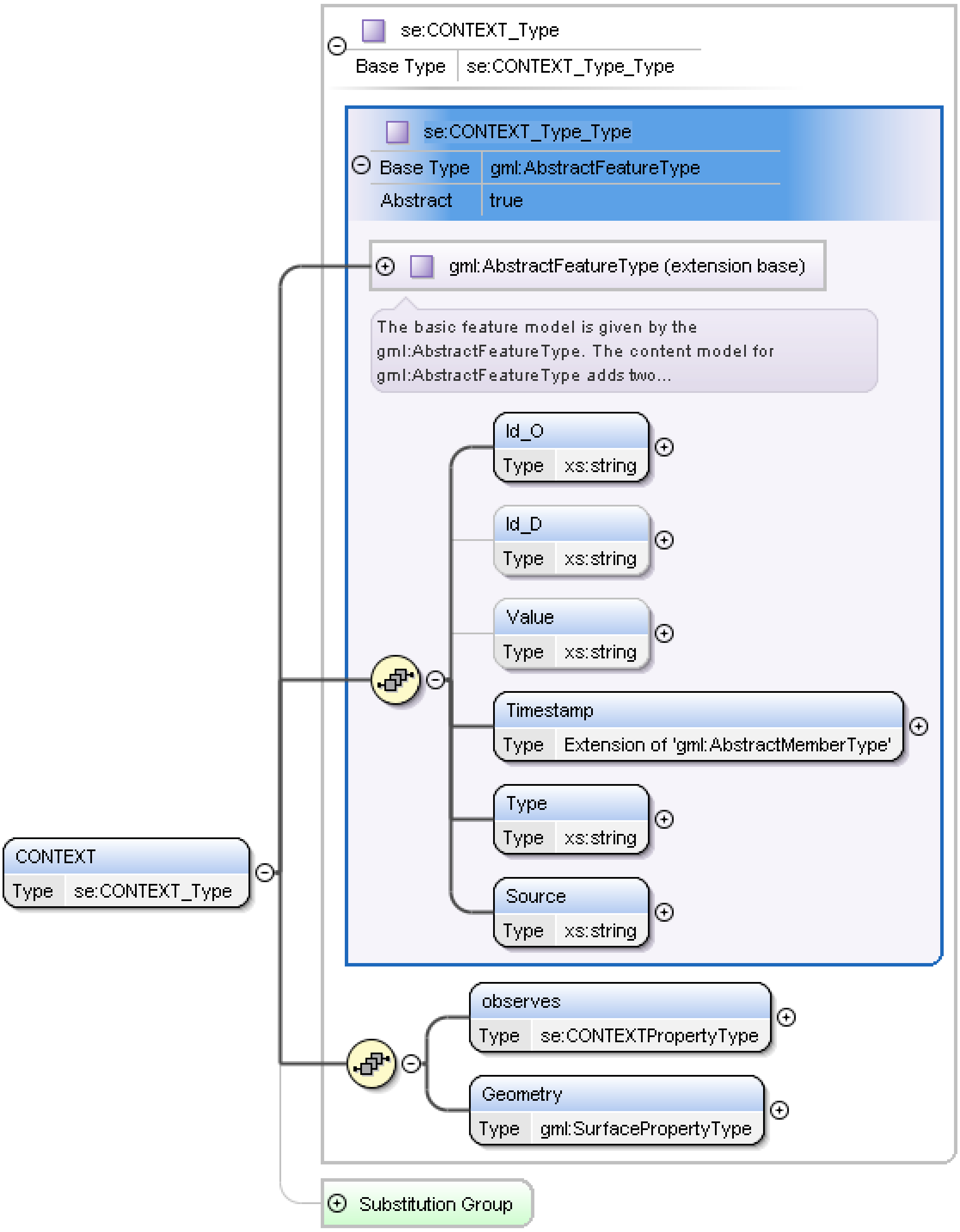
In this section, we specify the export data model, an XML Schema, of IDS-Geo. One expected result of the EHPS-Net project is that it should be possible to extract data from the IDS databases into rectangular datasets suited for statistical programs. However, even if the databases use the same relational database schema, data and software heterogeneities may occur due to different DBMS software programs being used (Yeung & Hall 2007). For example, an extraction program designed for one specific database may have to be edited before it can be used for other databases. Thus, standardised transfer formats, such as XML, may be used to overcome the heterogeneities. When using common and standardised exchange formats, we can also use standardised and open languages, such as Extensible Stylesheet Language Transformations (XSLT), to transform the data into the desired datasets. This will facilitate the sharing, reuse and development of the extraction and transformation programs.

XML is a platform-independent, text-based, free and standardised markup language that defines a set of rules on how to structure data. It is structured by elements, which have a function similar to that of attributes in a relational database. To specify the allowed structure and syntax of the XML document, an XML Schema is used (which is the XML correspondence to a database schema). Thus, we create such an XML Schema based on the IDS-Geo conceptual model. Figure 4 shows an extract from the XML Schema; in this extract, the schema specifies which elements are allowed in the CONTEXT element. Rules about these elements are specified as well. The full XML Schema is available on the journal website of *Historical Life Course Studies.*

For the geometric elements, we use the XML grammar Geography Markup Language (GML) (Lake et al. 2004). GML is an ISO standard for encoding geographic information developed by the Open Geospatial Consortium (OGC). It is primarily used for exchanging geographic data (Portele 2007). The geometries in GML are specified according to the GML Simple Features profile, which in turn is based on the OGC Simple Features geometry object model. However, the GML Simple Features profile is less restricted than OGC Simple Features and supports, e.g., 3D geometries (Portele 2007). There are also several versions of GML. GML version 3.2.1 is the current version, but many GIS software programs do not yet fully support this version. One reason for the lag is that 3.2.1 is able to structure data in a more complex way by using nested elements and abstract data types. The previous version, GML 2.1.2, is thus more widely supported because it is using a simpler and flatter data structure. Nevertheless, GML 3.2.1 is the version recommended by many geographic data sharing initiatives such as the INSPIRE Directive (EC 2009), and therefore we use this version to specify the geometries in IDS-Geo. One drawback with GML (and this holds true for XML as well) is that it generally requires larger storage space compared to other common GIS formats (Lu et al. 2007). This could be a problem when large datasets need to be transferred.

In the XML Schema (Figure 4), the (GML) geometry element is included directly in the observation element. Thus, each observation may have a geometry value, such as a polygon, described with GML elements. Moreover, if an online code list would exist, the elements Type, Relation and Timestamp could be controlled by it via hyperlinks. By doing so, we would be able to assure that only codes that have been agreed upon would be used. When describing dates, the XML data type date could be used. XML date is based on the ISO 8601 (ISO 2004) standard for exchanging date and time data (format YYYYMMDD or YYYY-MM-DD, with months and days being optional). Unfortunately, XML date does not fully comply with ISO 8601 because months and days are not optional, but mandatory, in this ISO-rule. This may cause problems when, e.g., only the year is known for an event. Therefore, we create both elements using the XML date type, as well as separate elements for days, months and years (as in IDS version 4).

Figure 4 *Extract of the IDS-Geo XML Schema showing the CONTEXT class.*



Source: The XML Schema is visualised in oXygen XML Editor.

Finally, the relationships between the classes can be described with the XML Linking Language (Xlink). Xlink is a language for creating and describing links within and between XML documents (W3C 2010). Another option is to include the referenced element directly in the referring element. This solution enhances the readability of the document and may facilitate the import to GIS programs. However, it duplicates data and, as discussed before, there is limited support for nested elements among several GIS programs. Therefore, we believe the best solution is to use Xlink.

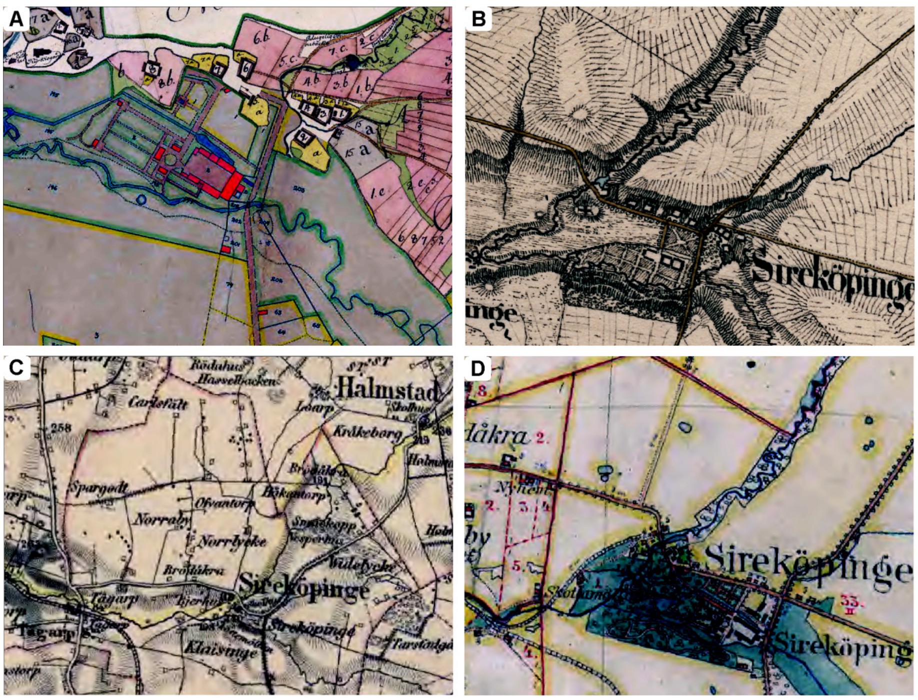
4 CASE STUDY

The aim of this case study is to: (1) implement the conceptual IDS-Geo model (Figure 3) for the geographically extended version of the SEDD version 3.1 IDS database, named SEDD IDS-Geo v.1; and (2) export the data from SEDD IDS-Geo v.1 into XML files compliant with the specified XML Schema (section 3.5).

4.1 MATERIAL

The Centre for Economic Demography (CED), Lund University, has developed the SEDD database during recent decades (Bengtsson et al. 2012), and it has been used extensively in research. SEDD contains demographic and economic information about all persons that have lived in five parishes, located in southern Sweden named Hög, Kävlinge, Kågeröd, Sireköpinge and Halmstad, from the 17th century onwards. The primary sources for SEDD are vital registers, annual poll-tax registers and continuous population registers. The database also includes extensive contextual information. A dataset in the IDS format are publicly available from version 3.1 of SEDD. The dataset covers the period 1813 to 1910 and contains observations of 79,656 individuals and 8,014 households. In a current project, we are establishing the SEDD IDS-Geo v.1 database, which is a geographically extended version of the SEDD version 3.1 IDS database. We have used approximately 60 historical maps from four map series: land surveyor maps 1757-1863 (LSM), military topographical survey map of Scania 1812-1820 (MTS), topographic maps 1860-1865 (TM), and economic maps 1910-1915 (EM) (Table 8, Figure 5). The historical maps have been scanned, geocoded and digitised. So far we have digitised approximately 900 property units to be stored according to IDS-Geo; in addition, around 3,000 buildings as well as a substantial amount of roads, railways, streams and wetlands have been digitised and stored in separate files outside IDS-Geo.

Figure 5 *Scanned historical maps covering parts of the Sireköpinge parish. (A) LSM; (B) MTS; (C) TM; (D) EM. (See Table 8)*



Source & Explanation: By using aerial photographs (from, for example, the 1940s) and modern maps, we have been able to geocode all the historical maps to the Swedish national spatial reference system SWEREF 99 (the Swedish realisation of the European ETRS 89 system). There is substantial ongoing work in linking the digitised geographic objects both with each other and with the individuals and contexts in the SEDD IDS database. In this case study, we only use property units for the Hög parish.

Table 8 *Digitised historical maps for the five parishes.*

Map series	Years	Digitised objects	Scale
Land Survey Maps (LSM)	1757 – 1863	Property units, buildings	ca. 1:5,000
Military Topographical survey (MTS)	1812 - 1820	Buildings, roads, streams, lakes	1:20,000
Topographic maps (TM)	1860 - 1865	Buildings, roads	1:100,000
Economic maps (EM)	1910 – 1915	Property units, buildings, railroad, roads, parish	1:20,000

Source: The 60 scanned and digitised maps come from the four map series.

4.2 DATABASE IMPLEMENTATION

We implemented the IDS-Geo conceptual model for the digitised property units of Hög parish. The SEDD IDS-Geo v.1 database is structured according to the low-level logical UML model shown in Figure 6. The model is specified for Microsoft SQL Server 2008 (which is the DBMS software that SEDD is using). For this implementation, the following rules were adopted:

1. The data type used for the Value field in the CONTEXT table does not support storage of geometric objects. Therefore, the geographic information is added to an external table (CONTEXT_GEOMETRY) supporting geometry.

2. The geometry data type is compliant with the OGC Simple Feature specification (ISO 19125-1). To present the geometry, the ISO standard OGC Well Known Binary (WKB) is used. WKB is a binary language used for storing and transferring vector geometries and their associated coordinate systems (ISO/IEC 2011). WKB is used for describing geometric objects in most of the commercial and open source database software programs.

3. The geometric representation uses the spatial reference system SWEREF 99, which is the Swedish realisation of the European Terrestrial Reference System 1989 (ETRS89) (LM, 2012). The map projection is UTM zone 33 N (i.e. we use SWEREF 99 TM).

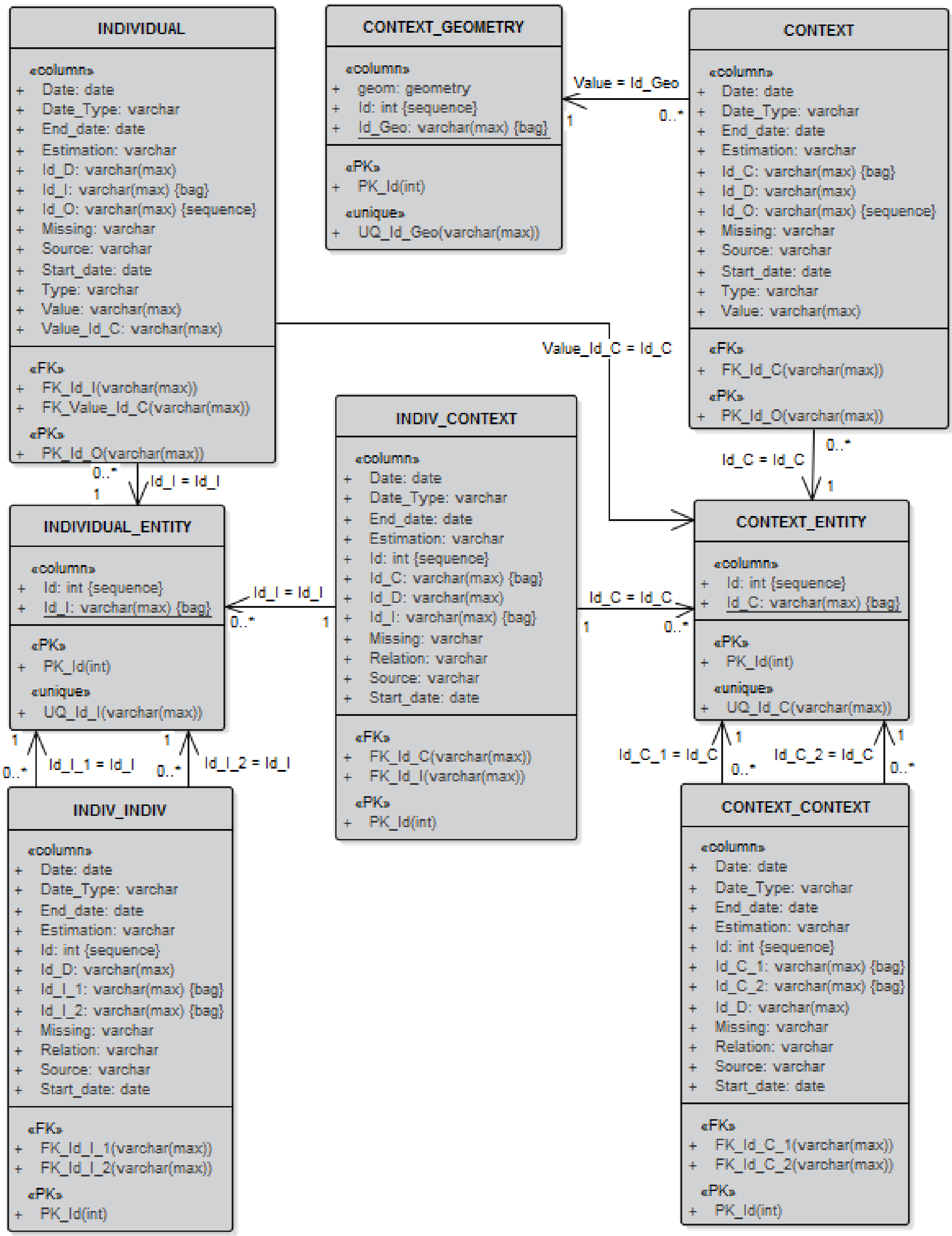
4. Each geographic object in CONTEXT_GEOMETRY has a unique identifier stored in the Geo_ID attribute, which can be linked to identifiers stored in the attribute Value in the CONTEXT table (which is the implementation of the CONTEXT class in the conceptual model Figure 3). The code geometry in the attribute Type is used to specify that the attribute Value contains such links.

5. The multiplicities of the tables are the following:
   a. One context may have several observations.
   b. One observation of a context may contain only one geometric object or collection.
   c. Several geometries can be observed for a context (at different temporal or spatial resolutions, or both).

   The result of the above points is that one context may have multiple geometric objects, and one geometric object may be used by multiple contexts.

6. Similar to the XML schema in section 3.5, we follow the ISO 8601 (ISO 2004) standard for exchange of date and time data (YYYYMMDD or YYYY-MM-DD, with months and days being optional). When there is a need to select certain months during a period, there exist methods such as getMonth to convert the ISO standardised dates into days, months and years.

Figure 6 *Logical model (low level) of the SEDD IDS-Geo v.1 database.*



To better illustrate the database implementation, Tables 9-11 provide examples of storing observations of three contexts (one property unit and two map sources). The objects are stored in the CONTEXT_ENTITY table, whereas their observations are stored in the CONTEXT table. The geometric representations in the CONTEXT_GEOMETRY table have links to the CONTEXT table through identifiers in the attributes Geo_ID (CONTEXT_GEOMETRY) and Value (CONTEXT) (see e.g., row 3 in Table 9 and row 1 in Table 11). It is also possible to link geographic objects and contexts in the CONTEXT_CONTEXT table, but this may produce slower calculations because more tables have to be joined in the queries. Based on the discussion in section 3.1, we choose not to use the IDS Type code Id_Polygon to link the observations to the geometric presentations because the linkage occurs within the IDS-Geo model. The purpose of Id_Polygon is to create links to external data outside IDS (and its name may limit the geometry types to polygons).

Table 9 *A CONTEXT_ENTITY table containing observations of the contexts.*

Id_O	Id_C	Source	Type	Value	Date	Start_date	End_date	Estimation
obs_34	pu_hog_1	SEDD	created	land reform	1804			
obs_3	pu_hog_1	lm_3	geometry	po_7492		1804	1840	intervalMinimum
obs_4	pu_hog_1	lm_3	geometry	po_7492		1804	1850	intervalMaximum
obs_20	pu_hog_1	poll-tax register	propertyFormation	subdivided	1850			
obs_5	pu_hog_1	em_2	geometry	po_32154		1850	1913	
obs_35	pu_hog_1	poll-tax register	propertyFormation	partitioning	1913			
obs_802	lm_3	lsm	mapCreation		1804			
obs_803	lm_3	lsm	mapName	Högs by	1804			
obs_804	em_2	em	mapCreation		1910			
obs_805	em_2	em	mapName	Hög	1910			

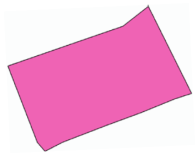
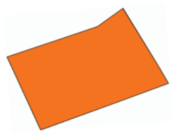
Explanation: The context pu_hog_1 is created during a land reform (obs_34). This context has three observations of two geometries: One polygon digitised from a land survey map 1804, and one polygon digitised from an economic map 1910-1915. Geometry po_7492 has two start and end date intervals; one minimum (obs_3) and one maximum (obs_4). That is, we know that geometry po_7492 exists at least until 1840, but it may exist up until 1850. For the second geometry (po_32154), there is no uncertainty interval. The start and end dates of geometry po_32154 are set to 1850 and 1913, respectively. The reason is that we know from the tax-poll registers that in 1850 the property unit po_hog_1 got subdivided and thus, changed its geometry (obs_20). Thereafter it got partitioned into several new and smaller units in 1913 (obs_35) and geometry po_32154 thus ceases to exist. The sources for the geometries are also stored (obs 802-5). Here, the source of geometry po_7492 is a land survey map named "Högs by" from 1804, whereas the source of geometry po_32154 is an economic map named "Hög" and created around 1910. The maps are part of the map series Land Survey Maps 1757-1863 (LSM) and Economic Maps 1910-1915 (EM). These map series will be described in the metadata about the database.

Table 10 *A CONTEXT_ENTITY table containing one property unit (pu_hog_1) and two map sources (lm_3 and em_2).*

Id	Id_C
12	pu_hog_1
31	lm_3
51	em_2

Explanation: lm_3 is the identifier for a Land survey map 1804, whereas em_2 is the identifier for an Economic map 1910-1915 (Figure 5).

Table 11 *CONTEXT_GEOMETRY table containing the geographic representations of the two observations of the property unit.*

id	geo_ID	geom	(Geometry view)
3	po_7492	(WKB Geometry)	
6	po_32154	(WKB Geometry)	

Explanation: WKB Geometry represents the geometry of a geographic object described using the OGC Well Known Binary data type (the actual WKB representations are hexadecimal strings with several thousands of characters and are thus not shown in the table). This data type also stores information about the spatial reference system used. The unique identifiers po_7492 and po_32154 enable the geometries to be linked to the observations in the CONTEXT table (Table 9). In this case study, the polygons of the property units are stored in one row each.

4.3 DATA XML EXPORT

Exports of the data are accomplished using XML documents valid to the XML Schema specified in section 3.5. The aim is to enable geographic data export from a standardised download service. In this case study, we transformed the data from the database into XML documents valid to the XML Schema. Figure 7 shows one geometry observation of the property unit described in Tables 9-11. In this XML example, the type observed is a geometry of property unit pu_hog_1, the start and end years of the observation are 1850 and 1913, and the GML geometry is a polygon with x and y coordinates specified in the posList element. In the Polygon element, the spatial reference system is specified using an SRID code (EPSG: 3008 is the code for SWEREF 99 TM). The source is the context em_2, which is the historical map used to digitise the property unit (when the source is stored as a context as well, a link could be established in the same case as with the <observes> element). In this example, the observation element is used as a root element. However, it is also possible to use the context as a root element with all of its observations inside it.

We have currently only exported few observations for property units in the Hög parish to XML documents. Based on these trials, we estimate that if each property unit on average has 10 observations (of which 1-2 are geometric observations), an XML document containing observations of our currently 900 digitised property units would have a file size of approximately 60 MB. If including data from the SEDD version 3.1 IDS dataset, we would have observations for around 80,000 individuals and 8,000 households (from the period 1813 to 1910). The file size of such dataset depends on the number of observations for each individual and household. For example, with an average of 10–30 observations per entity, the XML documents would, in uncompressed size, take up between 6 and 18 GB of storage space.

Figure 7 *Logical model (low level) of the SEDD IDS-Geo v.1 database.*

```
<featureMember>
    <CONTEXT xmlns="www.idsgeo.ed.lu.se" gml:id="obs_5">
        <Id_O>obs_5</Id_O>
        <observes>
            <CONTEXT_ENTITY gml:id="pu_hog_1">
                <Id_C>pu_hog_1</Id_C>
            </CONTEXT_ENTITY>
        </observes>
        <Source>em_2</Source>
        <Type>geometry</Type>
        <Timestamp>
                <Start_year>1850</Start_year>
                <End_year>1913</End_year>
        </TimeStamp>
        <Geometry>
            <gml:Polygon gml:id="po_32154" srsName="EPSG:3008">
                <gml:exterior>
                    <gml:LinearRing>
                        <gml:posList>
378525.06040099129,6184741.4450397845 378756.84782195603,6184236.7109536408
378518.65195828263,6184095.3461909313 378437.65232363169,6184278.1485594986
378252.16606974648,6184701.0288955392 378525.06040099129,6184741.4450397845
                        </gml:posList>
                    </gml:LinearRing>
                </gml:exterior>
            </gml:Polygon>
        </Geometry>
    </CONTEXT>
</featureMember>
```

5 DISCUSSION

Because the aim of the IDS structure is primarily to facilitate the exchange of historical demographic data, the main requirement for introducing geographic data in IDS-Geo is that database administrators can link individuals to spatial locations. For historical geographic data, such objects are, for example, buildings and property units. Other types of geographic data, such as topography and soil conditions, are not intended to be stored within IDS-Geo. They can, of course, still be used in longitudinal analyses as long as standard spatial reference systems are used.

One of the key questions in this paper was whether it was enough to use the external linkage to geographic data that IDS version 3 and 4 allows (using the Type code Id_Polygon in the CONTEXT table). Having external links to geographic databases have the benefits of simplicity; no geographic data need to be stored within IDS. Another advantage is that IDS data can potentially be linked to several geographic databases, but this requires standardized identifiers (e.g. modern addresses and building numbers). Historical geographic data, however, seldom have standardized identifiers. Although historical gazetteers exist (Southall, Mostern & Berman 2011), they are often at coarse scales. Therefore, when we want to assign detailed spatial locations to demographic data, we often have to create tailored geographic datasets. In those cases, we see benefits with storing the geographic data in the same structure as the demographic data, hence the need for a geographically extended version of IDS. Note that IDS supports storage of geographic point data, but in order to properly store building and property unit data, we also need to store line and polygon data; that is we have to add more geometric data types. Moreover, during the data export and exchange, having all of the data in the same data structure may facilitate the usability of the data, as well as the maintenance of the links between the demographic and geographic data.

One of the main purposes of EHPS-Net is to improve the data exchange needed for comparative longitudinal analyses. The following paragraphs discuss technical details on how such exchanges can be made easier for geographic data, using standardized formats.

In IDS-Geo, we make a distinction between the observation and the entity being observed, which may increase the query feasibility of the data. However, since it is important that IDS data can be easily converted to IDS-Geo, and vice versa, such distinction should not make a conversion more complicated. Therefore, when implementing the conceptual model in a database, the entity classes can be omitted and the observation classes can be used equivalent to the IDS version 4 CONTEXT and INDIVIDUAL tables. Users can therefore query a relational database implementation according to IDS version 4 (except for the added geometric data types). As for the IDS-Geo XML Schema, conversion should be simple as well, because there is a one-to-one relationship between the IDS tables and the XML Schema non-geographic elements.

The temporal representation of geographic data in IDS is an important issue. The generic nature of IDS allows for storing both object lifelines and event chronicles. However, for geographic data, there is a need to store intervals of not only the date, but also, the start and end dates (i.e., an interval of an interval). One solution is to add codes describing such intervals in the IDS Estimation attribute (e.g. intervalMinimum and intervalMaximum used in Tables 6 and 10).

IDS version 4 specifies a conceptual model for exchange of data. It also includes metadata explaining all of the variables and values used. Such model combined with metadata will solve two parts of the data heterogeneities that usually occur between data producers from several countries and domains, namely schematic/structural and semantic heterogeneity. The former is caused by the use of different data models to abstract the same real world concepts, whereas the latter is caused by different meanings of terms and concepts (Sheth & Larson 1990). Syntactical heterogeneities may, nonetheless, still occur. For example, different database implementations could use different data types and query languages, although they have the same database schema (Worboys & Duckham 2004). Software heterogeneities caused by different DBMS software can also be a problem. These heterogeneities complicate the process of developing extraction programs within EHPS-Net for exporting data from different databases. For instance, an extraction program designed for one specific database may have to be edited before it can be used for other databases. Nevertheless, these obstacles can be overcome if mostly common SQL queries will be used. Moreover, standards such as Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC) usually aid in data sharing among systems and databases. These standards, which are implemented in most systems, enable users to access databases from any application regardless of the DBMS (Yeung & Hall 2007).

We see benefits, however, when distributing data in standardised transfer formats, through standardised web services. For data containing geographic objects, the standard format is XML (using GML elements for geographic data) (ISO 2011). Although the web service solution may be more complex to implement in the beginning, it has several benefits. Using XML facilitates the sharing, combination and discovery of information. First, most of the data heterogeneities will be solved and data sharing between several domains becomes easier. Second, standardised and open languages for transforming and combining the data, such as XSLT, can be used. When using such languages, it will be easier to share and reuse the extraction programs used to transform the IDS data. If an online code list register was to be set up as well (see e.g., EC 2014), each code can then be referred to by Unified Resource Identifiers (URIs) in the XML documents. A drawback with XML formats is that they take up a relatively large disk space, which can be a problem when using big datasets. On the other hand, the case study reveals that the file size of the exported XML data may be reasonable to handle.

6 CONCLUSIONS

Including geographic data in IDS will improve longitudinal analyses by enabling individual-level spatial analysis; this inclusion is a good option when address data are not available. In our study, we introduced the IDS-Geo data structure, which is a slightly modified IDS model in which we added geometric data types to allow for storage of geometric representations of geographic objects. These modifications facilitate the linkages between individuals and (geocoded) geographic objects. Because the main aim of the IDS structure is to simplify the exchange of historical demographic data, we believe that only geographic data that can link individuals in IDS to spatial locations should be stored in IDS-Geo. Nevertheless, when standard addresses are available, they should be used instead of geographic data.

The IDS-Geo model was designed conceptually and an XML Schema (with GML elements specifying the geographic data) was created for the data export. Both of these models allow only geometries based on the OGC/ISO Simple Feature specification. We have also argued that using standardised exchange formats such as XML should aid in data sharing, as well as the development of extraction and transformation programs.

We implemented the conceptual IDS-Geo model in a case study using digitised property units stored in a geographically extended version of the Scanian Economic Demographic Database (SEDD). To fit into the IDS-Geo data structure we included an object lifeline representation of all the property units (based on the snapshot time representation of single historical maps and poll-tax registers). We tested also exporting the data according to the IDS-Geo XML Schema. The case study verifies that the IDS-Geo model is capable of handling geographic data that can be linked to demographic data. However, more research is required to test the usability of the model, using fully integrated individual level demographic and geographic data.

We have argued that it is easy to transform between the two models IDS and IDS-Geo. However, our intention in the long run is not to create a new geographic branch of IDS. Instead, we hope that this study could influence future development of the IDS structure.

ACKNOWLEDGEMENTS

The work with establishing the SEDD IDS-Geo v.1 database and service has been a collaboration between the Centre for Economic Demography (CED) and the Department of Earth and Ecosystem Analysis (ENES). Contributors to this work include Tommy Bengtsson and Clas Andersson from CED as well as Irene Rangel, Roger Groth, Mattias Spångmyr, Lena Arvidsson, Ali Mansourian and Daniel Persson from ENES. We also thank the two anonymous reviewers for their constructive comments that improved the quality of this paper. The work has been financed by the Swedish Research Council through the VR-project ESSENCE and through the Linnaeus grant for the Centre for Economic Demography. Some of the data for the database have been provided for free by several organisations, e.g., the County Administration Board in Scania, Stockholm University, Swedish Geological Survey, OpenStreetMap, Bing Maps and the Swedish National Heritage Board.

REFERENCES

Alter, G., Mandemakers, K. & Gutmann, M. (2009). Defining and Distributing Longitudinal Historical Data in a General Way Through an Intermediate Structure. *Historical Social Research,* 34 (3), 78-114.

Alter, G. & Mandemakers, K. (2012). The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 3, dated 12 July 2012. *Working paper* published on the EHPS collaboratory.

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4. *Historical Life Course Studies*, 1, 1-26.

Armstrong, M. P. (1988). Temporality in spatial databases. *Proceedings from GIS/LIS*, 88(2), 880-889. San Antonio, Texas.

Bengtsson, T., Dribe, M. & Svensson, P. (2012). *The Scanian Economic Demographic Database. Version 2.0 (Machine-readable database)*. Lund: Lund University, Centre for Economic Demography.
Berman, M. L. (2003). *A Data Model for Historical GIS: The CHGIS Time Series.* Cambridge, MA. Harvard Yenching Institute.
Cox, S. (Ed.). (2010). *OGC Abstract Specification: Geographic information — Observations and measurements* (OGC Abstract Specification).
Dam, P. (2013). *Integrating time and space in a digital-historical administrative atlas.* Unpublished manuscript.
European Commission (EC). (2009). Commission Regulation (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Services, *Official Journal of the European Union*, 52.
European Commission (EC). (2014). *INSPIRE code list register*, dated February 03, 2014.
European Historical Population Samples Network (EHPS-Net). (2009). *Proposal for an ESF Research Networking Programme – Call 2009.*
Fitch, C. A. & Ruggles, S. (2003). Building the national historical geographic information system. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 36(1), 41-51.
Gregory, I. & Southall, H. (2005). The Great Britain Historical GIS. *Historical Geography,* 33, 132-134.
Grenon, P. & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial cognition and computation*, 4(1), 69-104.
Herring, R. J. (Ed.). (2011). *OpenGIS Implementation Specification for Geographic Information - Simple Feature Access - Part 1: Common Architecture* (OpenGIS Implementation Standard).
INSPIRE Data Specifications Drafting Team. (2013). *D2.5: Generic Conceptual Model, Version 3.4rc3* (Framework Document).
ISO (2004). *ISO 8601:2004 Data elements and interchange formats -- Information interchange -- Representation of dates and times.* Geneva: ISO.
ISO (2011). *ISO 19156:2011 Geographic information - Observation and Measurements.* Geneva: ISO.
ISO/IEC (2011). *ISO/IEC 13249-3:2011 Standard, Information technology - Database languages - SQL multimedia and application packages - Part 3: Spatial.* Geneva: ISO/IEC.
ISO/IEC (2014). *ISO/IEC Directives, Part 1: Procedures for the Technical Work,* 11th edition. Geneva
Lake, R., Burggraf, D. S., Trninic, M., & Rae, L. (2004). *Geography Mark-Up Language: Foundation for the Geo-Web.* Chichester: John Wiley & Sons.
Lantmäteriet (LM). (2012). *Two-dimensional systems - SWEREF 99, projections.*
Lu, C. T., Dos Santos Jr, R. F., Sripada, L. N., & Kou, Y. (2007). Advances in GML for geospatial applications. *Geoinformatica*, 11(1), 131-157.
Portele, C. (Ed.). (2007). *OGC Implementation Specification 07-036: OpenGIS Geography Markup Language (GML) Encoding Standard* (OpenGIS Standard).
Sheth A. P. & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys* (CSUR), 22(3), 183-236.
Southall, H., Mostern, R. & Berman, M. L. (2011). On Historical Gazetteers. *International Journal of Humanities and Arts Computing,* 5 (2), 127-145. DOI: 10.3366/ijhac.2011.0028
Vanhaute, E. (2003). *Construction of a GIS for the territorial structure of Belgium.* Ghent. Ghent University.
Walker G., Taylor P., Cox S. & Sheahan, P. (2009). Water data transfer format (WDTF): Guiding principles, technical challenges and the future. In: R. Anderssen, R. Braddock, & L. Newham (Eds.). *Proceedings from 18th IMACS World Congress & MODSIM 2009 International Congress on Modelling and Simulation,* 4381–4387. Cairns: Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation.
Worboys, M., & Duckham, M. (2004). *GIS: A computing perspective.* Boca Raton: CRC Press.
Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1), 1-28.
World Wide Web Concortium (W3C). (2010). *XML Linking Language (XLink) Version 1.1: W3C Recommendation 06 May 2010.*
Yeung, A. & Hall, G. (2007). *Spatial database systems: Design, implementation and project management.* Dordrecht: Springer.
Yuan, M. & Hornsby, K. S. (2008). *Computation and visualization for understanding dynamics in geographic domains: a research agenda.* Boca Raton: CRC Press.