A Point-based Foundation for Statistics

Final report from the GEOSTAT 2 project

A Point-based Foundation for Statistics - Final report from the GEOSTAT 2 project

Eurostat ESSnet grant project

Grant Agreement number: 08143.2014.003-2014.736

Foreword

This report presents the main results from the GEOSTAT 2 project, an ESSnet grant project conducted between 2014 and 2016. The main objective of the project has been to propose a model for a point-based geocoding infrastructure for statistics based on geocoded address, building and dwelling registers. The outline of this model, as presented in this report, is based on national practices and specific national challenges, and also takes into consideration the EU context. The scope of the project has also included an evaluation of the Generic Statistical Business Process Model the (GSBPM) in terms of its fitness for purpose to describe the use of geospatial data in the production of statistics and to provide recommendations on the possible improvement of the GSBPM to better address geospatial data management in the statistical production process.

Chapter 1 to 5 present the methodological framework of the project and introduces and elaborates on the foundation of a point-based geocoding infrastructure at the conceptual level.

Chapter 6 elaborates on the operational aspects of a point-based geocoding infrastructure as to provide guidance to support NSIs in the process of setting up and maintaining such an infrastructure. A number of key tasks has been identified and mapped against the main phases of the Generic Statistical Business Process Model (GSBPM) in order to demonstrate how the geocoding infrastructure can be set up and used as an integral part of the general statistical production process. The key tasks proposed in Chapter 6 are illustrated with national use cases presented in annex 1.

Chapter 7 demonstrates how the GSBPM can help NSIs to mainstream the use of geospatial information in their statistical production process. It also offers suggestions on how the GSBPM could be extended to recognise better the use of geospatial information in the statistical production process. The GSBPM and its application to geospatial information have been approached through a set of workshops and exercises conducted jointly by the project consortium, as well as by project members and co-staff in their respective offices. The results from these exercises can be found in annex 2, 3 and 4.

The project has been coordinated by Statistics Sweden and supported by Eurostat. The project consortium comprised co-partners from six other countries. The project participants and contributors to this report have been:

- Marie Haldorson, Jerker Moström & Karin Hedeklint, Statistics Sweden
- Ingrid Kaminger, Statistics Austria
- Rina Tammisto, Statistics Finland
- Vincent Loonis, National Institute of Statistics and Economic Studies
- Ola Nordbeck & Erik Engelien, Statistics Norway
- Amelia Wardzińska-Sharif, Central Statistical Office of Poland
- Ana Santos, Statistics Poland

Stockholm, January 26, 2017

Content

1 Introduction

1.1 Background

As the world faces complex challenges to manage limited resources, there will be an increasing need for statistics to provide information on *how*, *when* and *where* specific resources should be allocated.[1] Only recently the United Nations Sustainable Development Goals (SDGs) have added a new set of targets that will need to be monitored through official statistics[2]. As a result, the need for accurate and reliable statistics has never been more apparent.

Through the link to administrative areas, geography has always formed part of statistics, but traditionally in a rather passive way, typically as a spatial container for the dissemination of statistical outputs. However, in order to respond to the *where* call, statistics need to be increasingly understood within the context of locations and temporal-spatial relationships. To this end, NSIs need to consider the location aspects throughout the statistical production process and improve their ability to integrate geospatial information in the collection, processing and dissemination of data. Only then, the full context of geography can be added to statistics.

The need to enhance the use of the spatial dimension during the collection, processing and dissemination of statistics has been acknowledged by the EU and its Member States for almost two decades. Drawing on the efforts made by Nordic countries, to develop harmonised gridded population data in the 1990s, Eurostat initiated the “TANDEM project” in 1999. The TANDEM project was a feasibility study towards a common geographical base for statistics across the European Union. The project argued that the system of NUTS areas of the EU, and administrative areas in general, was far from ideal, given the need for flexible and comparable geospatial statistics. Valuable geographical patterns are lost by the aggregation of data to the NUTS, and the comparability of territories grows increasingly weak. The project paved way for a harmonised small area output system for Europe and made a strong case for “a geographical base for statistics”.[3]

This work was resumed around the 2011 Population and Housing census, with the launch of the ESS long-term strategy to harmonise territorial statistics and to integrate geospatial information and statistics – GEOSTAT. In 2010 Eurostat launched the first ESSnet project “GEOSTAT 1 – representing census data in a European population grid dataset”. The aim of the project was to develop a vision and methodology for a European population grid dataset and, at a later stage, to produce the first version of a European dataset with population characteristics into a 1km² grid based on the Census 2011.[4] The final European GEOSTAT 2011 grid dataset was compiled from 28 national grid datasets, based on national census information, and gaps were filled with modelled information for only four

[1] UNECE 2016. In-depth review of developing geospatial information services based on official statistics. Note by the UK Office for National Statistics.

[2] http://unstats.un.org/sdgs/

[3] Eurostat 2002. TANDEM GIS I – A (feasibility) study towards a common geographical base for statistics across the European Union. Working papers and studies.

[4] EFGS/GEOSTAT 1 2012. GEOSTAT 1A – Representing Census data in a European population grid - Final report 2012; EFGS/GEOSTAT 1 2014 GEOSTAT 1B – Final report 2014.

countries in total, representing less than 2 million citizens or less than 1 percent of the population of the EU[5].

Besides successfully moving from a vision to practice by putting together the first ever pan-European, high resolution population dataset, and the first European geospatial statistics[6] product, the project concluded that "a point-based foundation" was needed to connect statistical information to spatial location. The project demonstrated that the "bottom-up" aggregation using a point-based spatial reference system for statistics resulted in far better data quality than a hybrid, let alone the disaggregation methods.

The successful results obtained in the GEOSTAT 1 project inspired the ESS to make further progress in achieving statistical-geospatial data integration along the entire statistical production process. In 2013 the DIMESA[7] mandated a Task Force on integration of geography and statistics to develop recommendations for improved information integration within the ESS. The Task Force endorsed the need for a point-based foundation for statistics, as proposed by the GEOSTAT 1 project. It also concluded that statistical-geospatial data integration should start at the level of individual statistical unit records. Consequently, when Eurostat launched a call for the second phase of the GEOSTAT project in 2014, the necessary geospatial infrastructure was in focus.

[5] EFGS/GEOSTAT 1 2012. GEOSTAT 1A – Representing Census data in a European population grid - Final report 2012; EFGS/GEOSTAT 1 2014: GEOSTAT 1B – Final report 2014.

[6] Location or extent are the main characteristics of geospatial statistics. Geospatial statistics is geocoded to small (in most cases below level 5) administrative or non-administrative geographies.

[7] DIMESA means Directors' Meeting on Regional, Spatial, Environment Statistics and Accounts.

2 Problem statement

2.1 The outset and identified needs for modernisation

Over the past 10 years, the demand for statistical information that is linked to locations has grown rapidly, with the demand coming from users in all sectors: government, commercial, not-for-profit, academia, and citizens. The growing demand for this location information, including geospatial statistics, has occurred because people, administrations, government bodies and other organisations want to gain insights into their populations or the topic of interest in increasingly greater temporal and spatial detail. An important facet of this insight relates to the location of the populations under consideration; in many instances the insights are specifically required at the local level. Moreover, rapid increases in the mobility of people, products and services trigger the need to develop greater insight into the complex dynamics within and between regions. Past assumptions about where people live, work, play or access services no longer hold in the modern world.[8]

Insights derived from the flows of people and business transactions between locations, as well as more traditional regional demographic and business statistics are being used in the growing trend towards “place-based decision-making”.

There are a number of drivers for a closer integration of statistics and geospatial information. One of the key areas of the ESS Vision 2020[9] is to harness new data sources comprising Big Data, administrative data and geospatial data. Using data from a range of sources, for multiple purposes, requires their integration into a common reference system of harmonised concepts, but also common locations and time. Therefore, users have not only increased their demand for location information but they also require simpler integration of data across various data sources used in their analyses.[10] Along with time, location and space are neutral and well-defined concepts and, hence, can be used to integrate data from a wide range of topics.

The requirement to develop an integrated statistical and geospatial solution for the 2021 round of population censuses has been repeatedly expressed in the UN context, e.g. by the UN Committee of Experts on Global Geospatial Information Management in its report from 2013 and the UN Expert Group on the Integration of Statistics and Geospatial Information (EG-ISGI).[11] Also in the ESS context the 2021 population and housing census is the first driver.

Within the National and International Statistical Systems there is a move towards an increased use of administrative data and registers for census purposes. In parallel, many countries have launched national geospatial strategies to geocode administrative records in order to support data linkage.

[8] UNECE 2016. Statistical and Geospatial Information – an Australian perspective on challenges and opportunities. Note by the Australian Bureau of Statistics.

[9] ESS Vision 2020. http://ec.europa.eu/eurostat/documents/7330775/7339647/ESS+vision+2020+brochure/4baffcaa-9469-4372-b1ea-40784ca1db62

[10] UNECE 2016. Statistical and Geospatial Information – an Australian perspective on challenges and opportunities. Note by the Australian Bureau of Statistics.

[11] UNECE 2016. In-depth review of developing geospatial information services based on official statistics. Note by the UK Office for National Statistics.

Official Statistics can take advantage of this increased data linkage through geocoded administrative data for the 2021 round of population censuses.

In response to the growing need to add the "where" dimension to statistics, NSIs have to be increasingly flexible to be able to deliver the range of statistics required. This flexibility can be achieved through the modernisation of official statistics which UNECE is facilitating by means of the Common Statistical Production Architecture and Generic Statistical Business Process Model (GSBPM). These are flexible frameworks that should allow geospatial information to be integrated into the statistical process without too much impact on the existing organisational structure.

The economic realities are such that the resources available to public information agencies, such as National NSIs and NMCAs (National Mapping and Cadastral Authorities), are becoming constrained with the increasing need to do more with less. These realities bring with them the challenge of having to maintain the quality and value of authoritative information in order for the users to maintain the level of trust in data which appears critical for evidence-based policies. It follows that, along with diminishing resources, it becomes increasingly more important that they are used effectively with a view to retaining their public value.[12]

Several international expert[13] groups have been discussion for a number of years on the best way to respond to the above challenges specifically from a geospatial perspective. The conclusion is that location needs to be tightly and fully integrated into the statistical production process using a point-based foundation for statistics. This point-based foundation for statistics would enable NSIs and the European Statistical System to answer the following needs:

- Increase relevance – by linking statistical information to specific locations, NSIs will enhance their ability to ground place-based decision-making policy in a spatial context.
- Increase efficiency – a more flexible production setup can provide a wide array of statistical outputs (in spatial and thematic terms) at low costs.
- Improve timeliness – by linking administrative data sources with authoritative location data, NSIs are better prepared to respond to the rapidly-growing users' needs.

2.2 Aim and scope of the project

The main objective of the GEOSTAT 2 project has been to propose a model for a point-based geocoding infrastructure for statistics based on geocoded address, building and dwelling registers. The outline of this model, as presented in this report, is based on national practices and specific national challenges, and also takes into consideration the EU context. The ESS vision of a fully geocoded census has been a priority of the proposed setup of a point-based geocoding infrastructure. However, the model should be suitable for official statistics and for other public and private data in the widest possible sense.

[12] UNECE 2016. United Nations initiative on Global Geospatial Information Management (UN GGIM) – All about connections. Note by United Nations initiative on Global Geospatial Information Management.

[13] Including the UN-GGIM Expert Group on the Integration of Statistics and Geospatial information, the GISCO working group, the European Forum for Geography and Statistics, and the ESS Task Force on the Integration of Statistics and Geospatial Information.

The scope of the project has also included an evaluation of the Generic Statistical Business Process Model the (GSBPM) in terms of its fitness for purpose to describe the use of geospatial data in the production of statistics and to provide recommendations on the possible improvement of the GSBPM to better address geospatial data management in the statistical production process.

In essence, the aim of the report is fourfold:

- To propose how a point-based geocoding infrastructure should be set up and maintained;
- To describe how such an infrastructure should be integrated into the statistical production architecture and processes in NSIs;
- To demonstrate how the GSBPM can help NSIs to mainstream the use of geospatial information in their statistical production process;
- To give advice on how the GSBPM could be extended to better recognise the use of geospatial information in the statistical production process.

3 Approach

The main approach to obtain an empirical input to the project has been an inventory of the existing national and sub-national spatial reference frameworks, and geocoding practises in the ESS Member States, candidate countries and potential candidates, conducted during the first year of the project. The inventory was carried out as a web-based questionnaire addressing NSIs. However, as the questionnaire contained a set of questions regarding the existence of geospatial data, respondents were urged to liaise with NMCAs or other authorities providing spatial data in their country before completing the questionnaire. The response rate was very good and the results from the survey proved to deliver a very exhaustive source of information. A full review of the results obtained in the survey can be found in the "Spatialising Statistics in the ESS" report[14].

Apart from the survey in question, a valuable input to the project has been collected through studies of a variety of reference materials, most notably reports produced by UN-GGIM, UNECE, Eurostat, previous GEOSTAT projects, different project reports and conference papers presenting national practises.

The GSBPM and its application to geospatial information have been approached through a set of workshops and exercises conducted jointly by the project consortium, as well as by project members and co-staff in their respective offices.

Last but not least, interactions with other related initiatives and colleagues in NSIs and NMCAs around the world have been important to elaborating the foundation on which the results are built. In the course of the project, the project members have participated in several working groups, such as the UN-GGIM Expert Group on the integration of statistical and geospatial information, the UN-GGIM: Europe Working Group B on data integration and the IAEG-SDG Working Group on geospatial information.

[14] EFGS/GEOSTAT 2 2016: Spatialising Statistics in the ESS – Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs.

4 The methodological framework of GEOSTAT 2

The GEOSTAT 2 project has pursued its work in a rich setting of initiatives on the global and European level to strengthen the integration of geography and statistics. Through the active involvement from the project consortium, the GEOSTAT 2 project has sought to enrich its own results, as well as to contribute to the progress of other initiatives. Most notable is the work conducted under the UN Global Geospatial Information Management, GGIM. Simultaneously with the launch of the GEOSTAT 2 project, UN-GGIM set up its European committee, which opened up for synergies between GEOSTAT 2 and the UN-GGIM: Europe Working Groups. The results obtained by the working groups (WG A Core Data and WG B Data Integration) have partially been incorporated in the results achieved within the GEOSTAT 2 project.

Another initiative with relevance for the project is the development of the Global Statistical Geospatial Framework (SGF)[15]. The emergence of the SGF has provided a solid context for the point-based geocoding infrastructure presented in this report. The GEOSTAT 2 results should be understood as the implementation guidance for NSIs, mainly as regards Principles 1, 2 and partially Principle 4 of the SGF (see Figure 3).

4.1 The Global Statistical Geospatial Framework

The international statistical and geospatial communities have recognised the challenge of a better integration of geospatial and statistical information, and have responded by establishing the United Nations Expert Group on the Integration of Statistical and Geospatial Information (UN EG-ISGI) to develop a Global Statistical Geospatial Framework (SGF). The SGF should act as a bridge between statistics and geospatial information, between NSIs and NMCAs, and between statistical and geospatial standards, methods, workflows and tools.

The SGF provides the international community with a common approach to connecting socio-economic and environmental data to specific locations, and improves the accessibility and usability of this geospatially-enabled data. Figure 1 below highlights the importance of location information as a tool integrating the following three domains: the society, the economy and the environment.

[15] United Nations Expert Group on the Integration of Statistical and Geospatial Information 2016. Background Document on Proposal for a Global Statistical Geospatial Framework (Advanced Draft as of 28/07/2016).

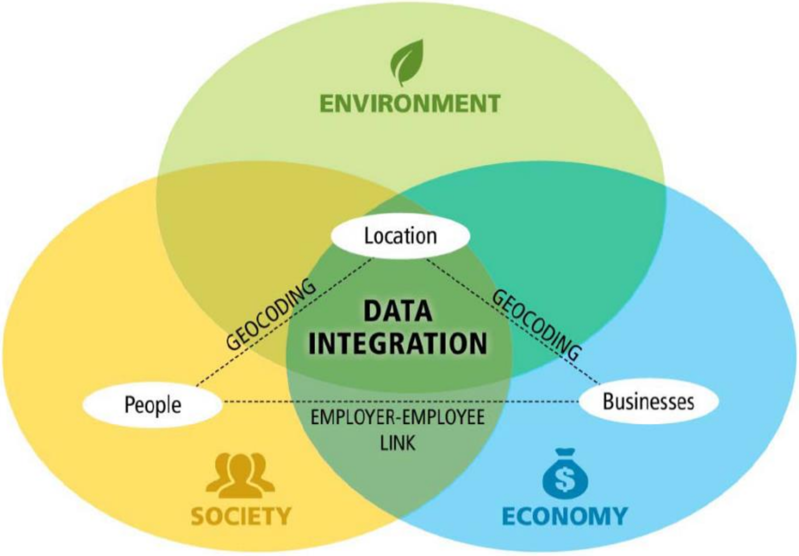


Figure 1: Location as a link between the society, the economy and the environment

In its first version, the SGF focuses on the socio-economic and environmental statistical data traditionally produced by NSIs (see Figure 1). The UN EG-ISGI will continue to develop the SGF and monitor its implementation with a review point in 2019. The intention of the Expert Group is for the SGF to be inclusive of all statistical and geospatial data, and to enable and encourage NSIs to look beyond traditional data sources and methods.

The SGF is a high-level framework that consists of five broad principles that are considered essential for integrating geospatial and statistical information (see the orange layers in Figure 2 below).

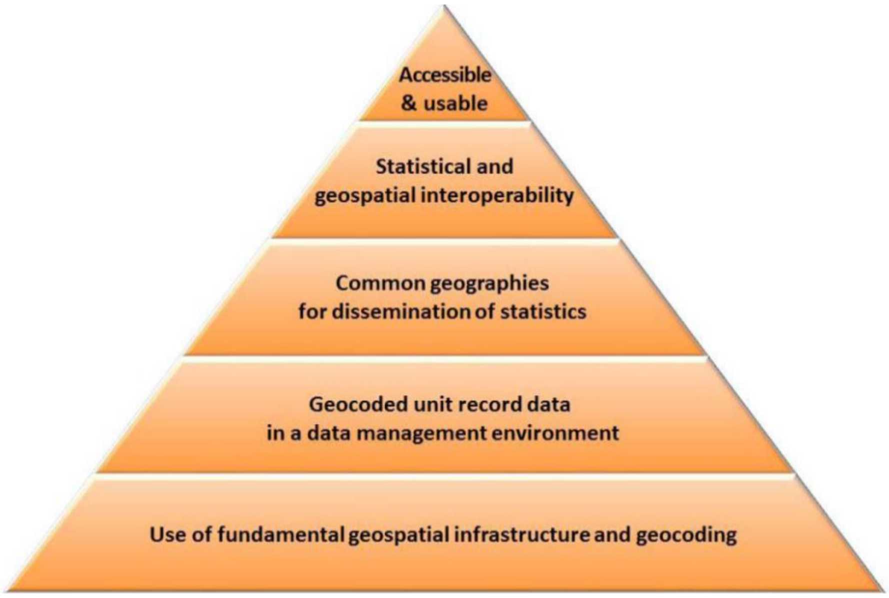

Accessible & usable
Statistical and geospatial interoperability
Common geographies for dissemination of statistics
Geocoded unit record data in a data management environment
Use of fundamental geospatial infrastructure and geocoding

Figure 2: The Global Statistical Geospatial Framework (SGF)

Each of the high-level principles in the SGF is defined by a set of goals and objectives, and is supported by international, regional and applicable domestic standards and best practices. These principles, along with the associated goals and objectives, are discussed below. The standards and best practices that will form the detailed guidance for countries implementing the SGF are still under consideration by the UN EG-ISGI, and will be brought to the UN Statistical Commission and UN-GGIM for consideration when finalised. The essence of the five principles is set out below but it is not the scope of this report to give a full introduction to the SGF.

Principle 1: Use of fundamental geospatial infrastructure and geocoding
The goal of this principle is to obtain a high quality, standardised physical address, property or building identifier, or other location description, in order to assign accurate coordinates and/or a small geographic area or standard grid reference to each statistical unit (i.e. at the micro-data level). Time and date stamping these locations will place the unit both in time and in space. The process of obtaining locations and geocodes should use relevant, fundamental geospatial data from the National Spatial Data Infrastructures or other nationally agreed sources. The geocoding of statistical units using point referencing is highly preferable when compared to merely associating statistical units with a geographic region (i.e. a polygon).

As the goal of the GEOSTAT 2 project has been to develop a draft model for a point-based geocoding infrastructure for statistics, Principle 1 can be considered substantially covered by the project (see Chapter 5 and 6).

Principle 2: Geocoded unit record data in a data management environment
The SGF recommends that the linkage of a geocode for each statistical unit record in a dataset (i.e. a person, household, business, building or parcel/unit of land) occurs within a data management environment. Persistent storage of high precision geocodes enables any geographic context to be applied when preparing the data to be released in the future (i.e. in aggregating data into a variety of larger geographic units or to adapt to changes in geographies over time). Moreover, geocodes can enable data linking processes that aim to integrate information of varying nature and sources by so called linked data techniques.

Many European NSIs have already developed, or are under way to develop, environments (“geography databases” or “key code databases”) for consistent linking of statistical information to location data. Principle 2 is partly covered in GEOSTAT 2 (see Chapter 5 and 6).

Principle 3: Common geographies for dissemination of statistics
To enable comparisons across datasets from different sources, the SGF recommends that a common set of geographies be used for the display, reporting and analysis of social, economic and environmental information. While the EG-ISGI recognises the importance of traditional statistical and administrative geographies, it also recommends NSIs to consider the benefits of gridded data. Gridded data can be both a rich source of information and a consistent geography for disseminating and integrating information. In Europe, the NUTS classification, functional territorial typologies such as the Degree of Urbanisation, the forthcoming TERCET regulation, and the INSPIRE specifications on statistical units including statistical grid systems provide already a solid basis to ensure comparable territorial data dissemination but further refinement is needed.

Principle 3 has not been in the scope of GEOSTAT 2 but GEOSTAT 1 has exhaustively studied the use of statistical grids as geographies for population and census information.

Principle 4: Statistical and geospatial interoperability – Data, Standards and Processes
Both the statistical and geospatial data communities operate their own general data models and metadata capabilities; however, these are often not universally applied. The statistical community uses the Generic Statistical Information Model (GSIM), the Statistical Data and Metadata Exchange (SDMX), and the Data Documentation Initiative (DDI) mechanisms. The geospatial community, on the other hand, makes use of the General Feature Model (GFM) and the ISO19115 metadata standard, plus a number of application specific standards.

Within the statistical community there is a need to build geospatial processes and standards into statistical business processes in a more consistent manner. In consequence, the EG-ISGI has recognised that a top-down approach is required with a view to incorporating geospatial frameworks, standards and processes more explicitly into the Common Statistical Production Architecture and its components. In particular, the Generic Statistical Business Process Model (GSBPM) needs to refer, to a larger extent, to the use of geospatial data and methods in the statistical production process, and in particular the data, standards and methods that are incorporated into the SGF.

GEOSTAT 2 will hopefully contribute to the development of Principle 4, both at the global and European level, as one of the main results of this project is to propose more viable ways of including geospatial processes into the GSBPM (see Chapter 7).

Principle 5: Accessible and usable geospatially-enabled statistics
This component of the SGF emphasises the need to identify or, where required, develop policies, standards and guidelines which support the release, access, analysis and visualisation of geospatially-enabled information. There is a wide range of legislative and operational issues that organisations need to be aware of when releasing and analysing information about people and businesses in a spatial context. One important aspect of this principle is to ensure that data can be accessed using safe mechanisms that not only protect privacy and confidentiality but also enable access to data in order to undertake various analyses that foster decision-making. Principle 5 has not been in the scope of GEOSTAT 2.

At the Sixth Session of the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM), held in August 2016, the principles of the Global Statistical Geospatial Framework were adopted. The background document[16] presented at the GGIM6 explains this framework in further detail.

When the GEOSTAT 2 project began, the SGF was still being elaborated. It has been very important to align the efforts in GEOSTAT 2 with the SGF, so that the project results would support the implementation of the SGF. As stated above, the GEOSTAT 2 results can provide guidance to NSIs as regards an approach to the implementation of Principles 1 and 2, and partially Principle 4 of the SGF. To this end, there will be a continuation in the next phase of GEOSTAT, the GEOSTAT 3 project. The main focus of GEOSTAT 3 will be to develop the European version of the Global SGF for the ESS, taking into account the existing conditions, initiatives and European and national frameworks.

[16] United Nations Expert Group on the Integration of Statistical and Geospatial Information 2016. Background Document on Proposal for a Global Statistical Geospatial Framework (Advanced Draft as of 28/07/2016).

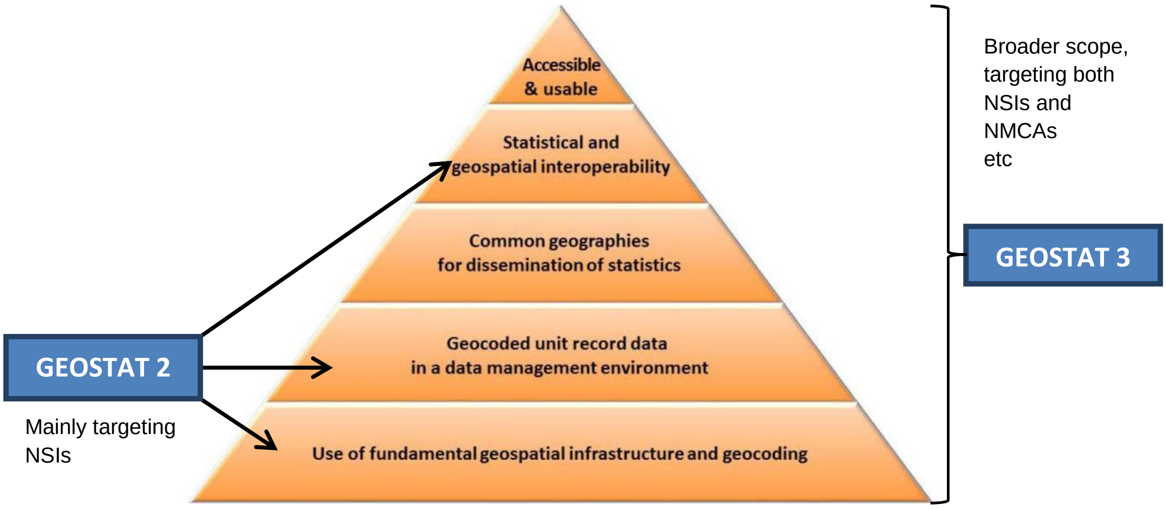

Accessible & usable
Statistical and geospatial interoperability
Common geographies for dissemination of statistics
Geocoded unit record data in a data management environment
Use of fundamental geospatial infrastructure and geocoding
GEOSTAT 2
Mainly targeting NSIs
GEOSTAT 3
Broader scope, targeting both NSIs and NMCAs etc

Figure 3: Illustration of the link between the GEOSTAT projects and the SGF.

4.2 The Generic Statistical Business Process Model (GSBPM)

The Generic Statistical Business Process Model (GSBPM) is a tool to understand and describe business processes. UNECE also provides other supportive models, which include the Generic Statistical Information Model (GSIM[17]), the Common Statistical Production Architecture (CSPA[18]) and the Generic Activity Model for Statistical Organisations (GAMSO[19]).

The GSBPM documentation lists the following benefits of using the GSBPM: "The GSBPM describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. The GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement."[20]

A better understanding of the relation between geospatial data and statistics is a key element to effectively mainstream the use of geospatial information in the statistical production process. It is the belief of the project that the GSBPM may help NSIs to design cross-product infrastructures for geospatial information, i.e. the same infrastructure might be used for several statistical production processes.

The documentation of the GSBPM includes descriptions of how to use the model in the statistical production process. However, the geospatial dimension is completely absent. The GEOSTAT 2 project therefore decided to evaluate the GSBPM[21] in terms of its fitness for purpose, in order to describe the use of geospatial data in the production of statistics and provide recommendations on the

[17] UNECE 2013: http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification
[18] UNECE 2015: http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.5
[19] UNECE 2015: http://www1.unece.org/stat/platform/display/GAMSO/GAMSO+v1.0
[20] UNECE 2013: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0
[21] The other related models (GSIM, CSPA and GAMSO) were not evaluated. However, they seem to be relevant at various levels of producing geospatially-related statistics, e.g. data modelling, service-based statistical production and management of geospatial data among other essential data sets.

possible improvement of the GSBPM to better address geospatial data management in the statistical production process. The result of this work is presented in Chapter 7.

The GSBPM and the SGF are complementary rather than competing concepts. Connecting the GSBPM and the SGF represents a process-oriented dynamic perspective on geospatial statistics as an integrated part of the statistical production process. As such, incorporating the principles of the SGF into the GSBPM is a step towards its concrete implementation.

4.3 A generic geospatial-statistical data model for the production and dissemination of statistics

In order to make recommendations, particularly on the implementation of a point-based geocoding infrastructure for statistics in the ESS, the full scope of geospatial information in statistical production first needs to be recognised. From a statistical production perspective, it is important to distinguish information needed as the infrastructure to geocode data from geospatial information needed to create statistical content. The GEOSTAT 2 project suggests that the different roles of geospatial information in statistical production can be described by means of a generic geospatial-statistical data model.

One of the main conclusions of the ESS Task Force on the integration of statistical and geospatial information was that countries should agree on a unique official reference data per country for geocoding statistics and producing geospatial statistics with clear ownership, defined scales and attributes taking into account statistical requirements. Both administrative data sources and survey information should be geocoded to the same reference system. In general terms, all Member States should make use of these geospatial reference data, mandatory for all public stakeholders at all government and administration levels, and for all public data and administrative tasks.[22]

The Task Force concluded that the geospatial reference data for statistics should be understood in the widest sense as "***a consistent framework of geospatial data needed to produce and disseminate statistics"***. The framework elaborated by the Task Force has a broader scope than the mere geocoding of statistical or administrative information. It also encompasses geospatial information on which statistical content can be built and disseminated. Hence, the geospatial reference framework elaborated by the Task Force will hereafter be referred to as ***a generic geospatial-statistical data model for the production and dissemination of statistics***. The GEOSTAT 2 project believes that this is a more accurate description of the concept. According to the geospatial-statistical data model, three different tiers of information can be identified, reflecting that geospatial information can be used either as infrastructure data or as data to create statistical content or as a combination of both purposes.

- Tier 1 comprises geospatial information used exclusively for the purpose of geocoding, geographically representing or disseminating statistical or administrative information. Such geospatial information is instrumental in the sense that is does not have any intrinsic value to statistics. It simply does not make sense to use this kind of information in statistical production unless it is linked to other information. Examples of information in Tier 1 include

[22] Eurostat 2015. Report from the Task Force on the Integration of Statistical and Geospatial Information 2015.

address data, census enumeration districts, postal code areas, statistical grids or other statistical or administrative geographies.

- Tier 2 comprises geospatial information which is used both to geocode, geographically represent or disseminate statistical or administrative information *and* to create statistical content. Typical information found in Tier 2 includes building information, cadastral parcels and transport networks but also new data sources such as traffic sensor information. The mixed purpose can be illustrated by building location data which is a key dataset to geospatially enable census data, but it is also used to calculate information on building density and building footprints, to assess the degree of urbanisation, etc. The geospatial object of a building has a value for statistics that goes beyond its location.
- Tier 3 comprises geospatial information which is used only to produce statistical content. Despite its geospatial component, this category of information cannot be used directly to geocode statistical or administrative data. As such, information in Tier 3 can be regarded as *complementary* to, and *independent from,* information in Tiers 1 and 2. Some examples of data found in Tier 3 include DEMs, land use or land cover maps, topographic data, ortho-photo or satellite imagery, or other products derived from earth observation data. In theory, Tier 3 data can be any geospatial data suitable to create statistical content. However, in order to guarantee the adequate quality of statistics, Tier 3 data should encompass authoritative datasets or other properly-documented and trusted datasets. Typically, data in Tier 3 needs to be combined with data from tier 1 or 2 in order to be transformed into statistical information. The calculation of land area within a NUTS region can serve as an accurate example. In this case authoritative data on land mass and topographic maps data (Tier 3) is combined with a dataset containing NUTS regions (Tier 1).

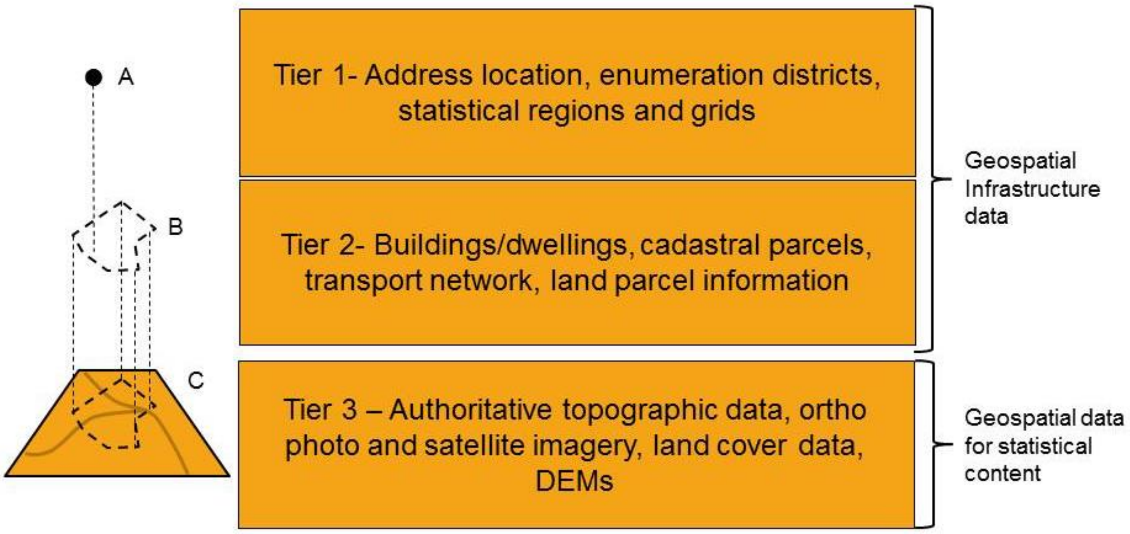

A
B
C
Tier 1- Address location, enumeration districts, statistical regions and grids
Tier 2- Buildings/dwellings, cadastral parcels, transport network, land parcel information
Tier 3 – Authoritative topographic data, ortho photo and satellite imagery, land cover data, DEMs
Geospatial Infrastructure data
Geospatial data for statistical content

Figure 4: Tiers of information in the generic geospatial-statistical data model for the production and dissemination of statistics. A workplace geocoded to an address location (A) can be linked to a cadastral parcel (B) from which land use can be assessed by combining the parcel with a land use map (C). The more consistent the system, the more opportunities for flexible linking of data.

In conclusion Tier 1, 2 and 3 in its totality represent the geospatial data that are required to geo-enable all relevant statistical information from enumerations, surveys, administrative and alternative data sources at the unit record and aggregate level throughout the statistical production process. Tier 1 and 2 are of a more fundamental nature and are therefore defined as the geospatial infrastructure data for statistics with the main purpose to geocode, spatially integrate, disseminate or represent statistics (e.g. on maps). As GEOSTAT 2 mainly deals with the conditions for geocoding statistics the rest of this report will therefore focus on Tier 1 and 2. The full statistical geospatial data model will be discussed again as part of the follow-up project GEOSTAT 3.

The categorisation of geospatial information outlined above is different from the traditional or emerging categorisations of geospatial information, such as topographic data, core data, fundamental data or reference data. As an example, the working group on core data of UN-GGIM: Europe has put forward 14 INSPIRE themes as core data that comprise many of the themes of tier 1, 2, and 3 of this categorisation[23].

Although both categorisations have their specific applications, it should be avoided that their applications create confusion and erect communication barriers. This will be followed up together with UN-GGIM and GEOSTAT 3.

[23] UN-GGIM Europe 2016. Core Data Scope. Working Group A - First Deliverable of Task 1.a.

5 A point-based geocoding infrastructure for statistics

Building on the generic geospatial-statistical data model for statistics, as outlined above, this chapter introduces and elaborates on the foundation of a point-based geocoding infrastructure at the conceptual level. The aim is not only to facilitate a common understanding of the characteristics of this geocoding infrastructure but also to outline the benefits, limitations and challenges associated with a point-based infrastructure specifically for statistics. Generic considerations with regard to the infrastructure set-up and use are discussed, and some recommendations are presented. However, more operational aspects related to the infrastructure are further explored in chapter 6.

5.1 What is a point-based geocoding infrastructure?

In a fundamental sense, a point-based geocoding infrastructure for statistics can be understood as a production setting where a record holding X, Y (and Z) coordinates of a location, along with a unique identifier (Id), can be linked to a record with statistical or administrative data which belongs to this point. This process is called “geocoding” of statistics or other data. The actual purpose of the point-based approach is to assign a single coordinate location to each unit record. The term “point-based” should be understood in contrast to “area-based” which appears in traditional surveys and censuses where the population surveyed is assigned to a fixed output area, such as an enumeration district. It should be stressed that the proposed shift from an area-based to a point-based approach, as described here, only refers to the geocoding infrastructure itself and hence to the collection and processing of statistics. The area-based approach is, and will continue to be, the primary method for the dissemination of statistics.

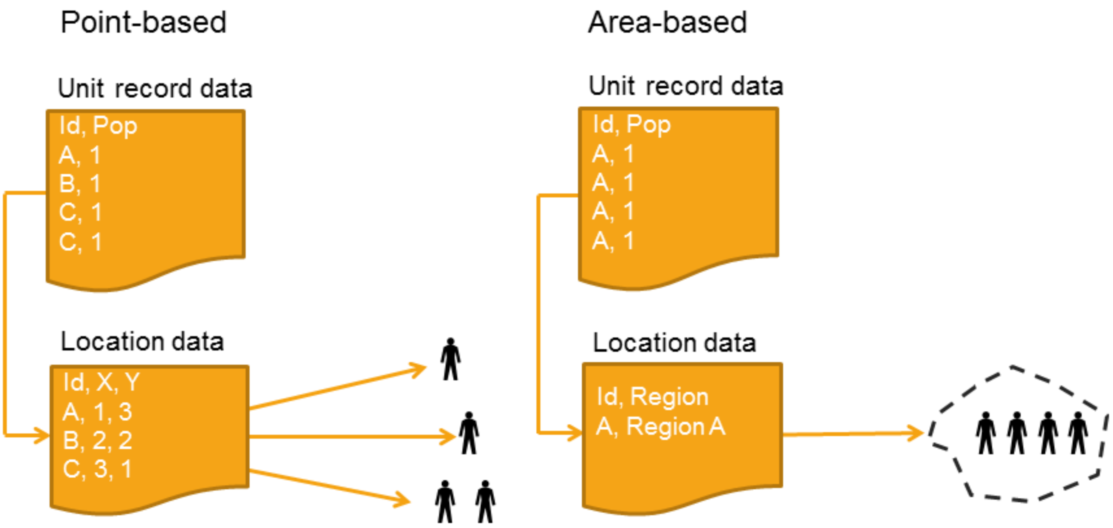


Figure 5: The conceptual difference between point-based and area-based geocoding infrastructures.

In Figure 5 above, the conceptual differences between a point-based and an area-based geocoding infrastructure are illustrated. In both cases there is a record with statistical data comprising four individuals, and a corresponding record containing location data. In the point-based approach, shown on the left, each individual in the statistical data record is linked to a unique dwelling location which has not been aggregated and is spatially represented by three different point locations. Two individuals have been assigned the same location as they are linked to the same dwelling. In the area-based approach on the right, all four individuals are linked to the same spatial object (Block A), as the area-based approach does not support spatial discrimination of their individual dwelling locations.

The underlying assumption of the GEOSTAT 2 project is that a point-based geocoding infrastructure is far more flexible in terms of production and maintenance than a traditional area-based infrastructure with fixed output areas, such as enumeration districts or other small areas. The point-based infrastructure is basically a system to integrate data in order to better exploit the spatial dimension of statistics. As such, it does not presuppose a specific mode of data collection. The point-based approach can be implemented in the context of traditional Census data collection, as well as in the administrative data-based context. However, as one of the key goals of the ESS Vision 2020 is to better exploit new data sources for statistics NSIs should opt for a point-based infrastructure based on authoritative location data, along with use of administrative data as this also allows easier integration with other data sources sharing the same location.

5.2 Constraints and challenges

In theory, the point-based approach is a simple concept. However, in practice it may encompass several challenges. The first, and by far most important, precondition for the successful implementation of a point-based approach is access to high quality location data, such as address information or building location data. According to the survey conducted within the GEOSTAT 2 project[24], the conditions regarding access to, and use of, these data sources vary strongly among the ESS countries. Data sources qualifying for high quality point-location data exist in the majority of the ESS countries. However this does not necessary imply that they are used as a point-based infrastructure to geocode statistics. There are several reasons behind this:

- The geospatial data coverage may be incomplete as address or building location data exists only in urban areas or in certain regions of the country
- Access to geospatial data is restricted either for legal or by financial reasons. Data can simply be too expensive.
- The quality of geospatial data is too poor; it is outdated with regard to the statistics.
- No consistent legal, technical and organisational framework for official geospatial data exists. The role of official data is crucial as statistical institutes needs to be able to rely on a long-term data provision strategy.

Altogether these conditions put strains on the development of a point-based infrastructure for statistics. At the same time area-based census frameworks have been successfully used for decades in many countries. Changing the approach and systematically geocoding all census information to location points requires substantial investments and can only be expected if the basic conditions are sound and safe.

The concept of a point-based geocoding infrastructure may not be fully embraced also for other reasons. Although high quality location data exists, and is both sound and accessible, legal restrictions on collecting and storing non-aggregated population data may foil ambitions to establish

[24] EFGS/GEOSTAT 2 2016. Spatialising statistics in the ESS. Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs.

infrastructures to geocode individuals to the single coordinate level, even in internal production databases of NSIs.

5.3 Characteristics of a point-based geocoding infrastructure

The characteristics of a point-based geocoding infrastructure encompass the following three generic principles:

- Use of high quality point-based location data, regularly updated with time stamps
- Geocoding of statistical unit, and related statistical information, at unit record level
- Use of standardised identifiers/geocodes to link unit record data with location data

High quality point-based location data should be understood as geospatial information that accurately represents the geographic location of a given phenomenon. The accurate point-based representation of an individual or a dwelling typically requires the use of a geocoded address or building data. Depending on various traditions throughout Europe, the rationale for choosing one of the categories over another may vary between countries. The GEOSTAT 2 project has concluded that it is of less importance whether the geocoding infrastructure is built on address data, building data or cadastral parcel information, as long as it can produce harmonised output with equal quality cross countries.

Hence, the choice of location data objects should rather be guided by the principles of authoritativeness and maturity of the location data, as well as by the potential for long-term temporal maintenance. According to the GEOSTAT 2 survey temporal accuracy and well-managed maintenance policies are rated even higher than the spatial accuracy of location data.[25] Yet, the topological and geometrical accuracy requirements may play an important role. As an example, the reference points need to be on the correct side of a street, and fall into the correct postal code areas, enumeration areas, electoral areas, etc.

In some Member States, location data frameworks comprise integrated combinations of address information, building or dwelling data and cadastral parcels. Ideally, these objects are consistently and hierarchically linked to each other, enabling a flexible choice of the location data objects to be used, depending on the purpose of the task and the scope of output data.

UN-GGIM: Europe Working Group A has recognised address, building and cadastral parcel information as core data.[26] Being part of the core data concept means that these data sources will gain higher priority in the work towards a more harmonised production and distribution of pan-European geospatial data.

Geocoding of statistical information at the unit record level means that each statistical unit record included in a dataset should be assigned a high accuracy geocode, i.e. without previous data aggregation or grouping. The geocodes assigned to each statistical unit record need to match the address codes or building codes found in the corresponding location data framework. Also, records

[25] EFGS/GEOSTAT 2 2016. Spatialising statistics in the ESS. Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs.

[26] UN GGIM: Europe 2016. Core Data Scope. Working Group A – First Deliverable of Task 1.a.

from different statistical data collections or administrative data sources need to be assigned to the same location code to allow for cross-domain data integration.

The use of standardised identifiers/geocodes to connect statistical information with location data means that the geocodes or identifiers used should be based on a nationally and officially agreed coding system that is unambiguous. The latter requirement means that postal addresses are only the second best choice as they are prone to inconsistencies, incompleteness, updates and duplicates. However not all countries have implemented their national standards for properly managing address information.

Finally, it is worth noting that, in addition to the three spatial dimensions of the geocoding infrastructure, time as the fourth dimension is equally important to consistently and unambiguously geocode or georeference statistical unit record data. All features making up the infrastructure need to be time-aware and have a start- and end-date, or at least very good metadata to know when and how the geocode was derived.

5.4 Approaches for setting up a point-based geocoding infrastructure

Though the concept of a point-based geocoding infrastructure is generic and the underlying principles will be the same cross countries, the production setting may differ between individual countries due to various traditions in data collection and governance between the institutions involved.

On the basis of the results obtained in the survey conducted in the GEOSTAT 2 project, three different approaches can be broadly identified among those countries that already have a point-based infrastructure in place. As further elaborated below, each approach has its own set of benefits and challenges in terms of performance and maintenance.

1. “In-house” – both location data and statistical data are collected and managed completely within the NSI

The “in-house” approach emerges when, for different reasons, third party authoritative address or building/dwelling information has not been accessible or fit-for-purpose. Hence, the NSI has made a strategic decision to build its own location data in order to set up the infrastructure needed for a fully-geocoded census. The benefit of this approach is that the NSI has full control of the entire data collection chain, including the coding systems (identifiers to link a location at the unit record level) and quality of both location data and statistical data. As the entire chain of data is in-house, the NSI does not have to deal with inconsistency problems resulting from changes to the address data standards, etc., at least as long as no third-party data, e.g. from tax authorities, needs to be geocoded.

High data collection and maintenance costs constitute the drawback of this approach, as the NSI needs to bear the costs on its own. In addition, this approach may also imply the lack of synergies regarding location data. Due to the confidentiality legislation, the NSI may not be allowed to share location data with other public institutions and vice versa. As a result other data producers may use different geocoding data that are not consistent with each other. This limits the use of harmonised geocodes in records within public administrations, and eventually hampers data integration and/or the coherence for these datasets which are not produced by the NSI.

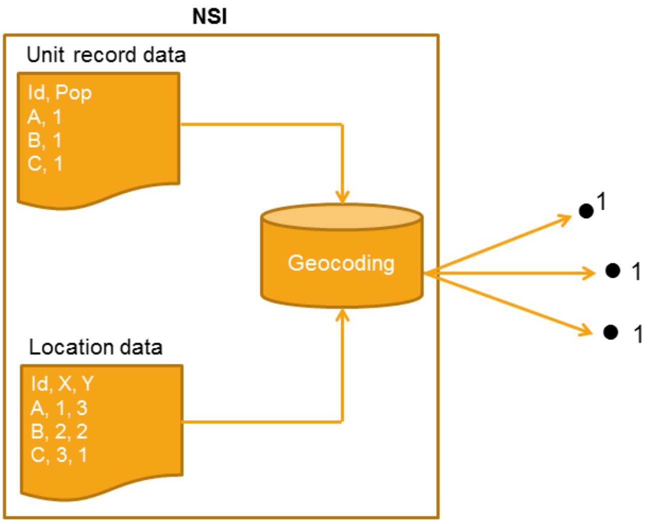


Figure 6: A conceptual illustration of the "in-house" approach

2. The "hybrid" – location data is collected and managed outside the NSI and statistical data within

The "hybrid" approach is quite common and can be found in countries with a well-developed framework of authoritative location data in combination with traditional census data collection settings. Typically, the responsibility for the collection and maintenance of location data rests with the NMCA or with a consortium consisting of the NMCA and the municipalities.

Shared costs of the location data collection and maintenance are the obvious benefit of the "hybrid" approach. Potential synergies in the use of location data constitute another benefit. If a framework of authoritative location data exists, other public administrations are most probably using the same data, which provides a ripe environment for an increased use of geocoded administrative data as the source for statistics and better data integration.

The challenge of this approach is that NSIs have to consider themselves one of many stakeholders, in terms of the national location data policy. Policies regarding the maintenance of geospatial data in the "outer world" will have a strong impact on the internal statistical data maintenance strategies. NSIs need to spend more time on monitoring and interfering with policies regarding the location data collection in order to safeguard consistent coding. Typically, some kind of a formal agreement is required between the NSI and the producers of location data to ensure long-term data access. In case no such agreements are in place, there is a risk that changing business models or priorities among data providers will endanger timely access to geospatial information.

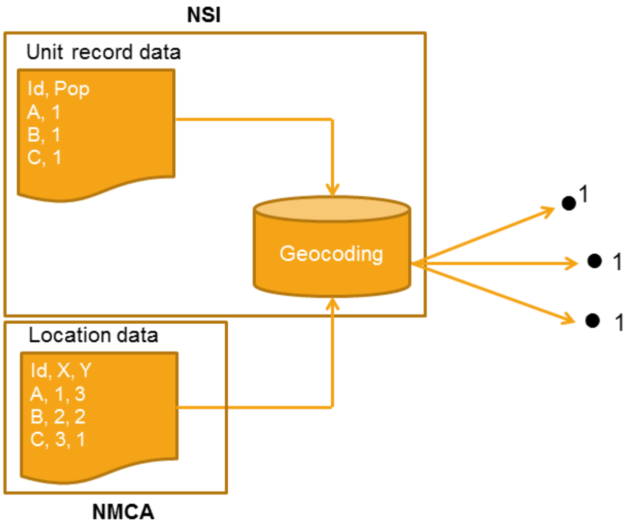


Figure 7: A conceptual illustration of the "hybrid" approach

3. The "data broker" – both location data and statistical data are collected and managed outside the NSI

"Data broker" implies that the main approach of the NSI is to put together data collected by others. The NSI obtains location data from NMCAs or other providers of geospatial information whereas data for statistical content are mainly obtained from other public administrations (tax administration, etc). Typically, access to administrative data is regulated in the statistical legislation, whereas access to location data is governed by separate agreements.

As in the case of the "hybrid" approach, shared costs of data collection and maintenance are the obvious benefit of this approach is. In the "data broker" approach, shared costs also apply to data to be processed into statistical information, e.g. administrative sources. There is also a great flexibility in terms of production as data is typically based on administrative sources, updated monthly, weekly or even daily. The use of authoritative location data with officially agreed geocodes forms the basis for registers/records in basically all public administrations and ensures consistency of all data, which creates almost unlimited data linkage opportunities

The downside of the "data-broker" approach is no or little direct control of data collection. As the collection of administrative data is not conducted primarily for statistical purposes, NSIs also have to deal with the fact that data may be structured in a way that requires substantial restructuring procedures to make it fit for statistical purposes. In this type of production mode, NSIs typically have to spend a considerable amount of time and resources to try to influence data collection policies within other administrative bodies. This can be a cumbersome task as there may be legislative strains on the collection of data which is not needed to perform the administrative task, for which it is collected, but which proves crucial for statistical purposes. In some cases administrative data can come geocoded from the custodian. This can pose problems in case the geocoding method is unknown or different to the methods used by NSIs.

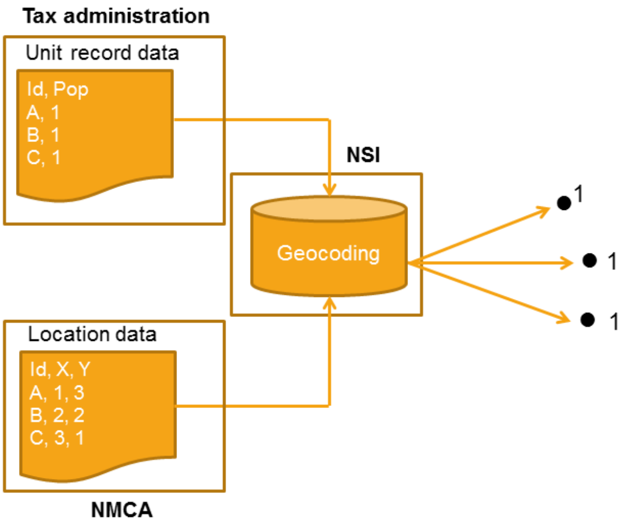


Figure 8: A conceptual illustration of the “data broker” approach.

In principle, all NSIs in the ESS can be assigned to some of the approaches described above. Yet, the approaches are generalised as the production setting may comprise elements from different approaches. Mixed use cases can also be identified where different approaches apply to the same NSI, depending on the statistical domain (e.g. social or business statistics), the task to be conducted and the information and specific infrastructure used for this purpose. Nevertheless, the three different approaches conceptualise some fundamental differences that need to be understood in order to provide the relevant guidance on how to build and maintain a point-based geocoding infrastructure.

5.5 Can and should all kinds of information be geocoded to point location?

Without questioning the fundamental benefits of a point-based geocoding infrastructure, a relevant issue to address is whether or not all information can or should be geocoded to point location. Another relevant question is whether the official address and building location data is sufficient to build a complete point-based geocoding infrastructure. To what extent is there a need to geocode data to the level of point coordinates where address or building location data may not be appropriate or sufficient sources?

According to the respondents to the GEOSTAT 2 survey, with some few exceptions, a well-managed official location data infrastructure with address and/or building information as its backbone fits the needs for geocoding most of the information found in NSIs or in other public administrations. The exceptions mentioned by the respondents can be divided into two categories:

- Cases where the data subject does not have a clear or easily discernible location (e.g. mobile or internet transactions, population in mobile households or certain types of transport statistics).
- Cases where the use of address or building locations will produce a non-appropriate spatial representation of the phenomena.

Some examples of the latter category include agricultural holdings and discharge points for water and pollution from industries. Agricultural holdings may be inappropriately geocoded by using address information when the address location refers to the dwelling of the farmer, rather than to the farm site of the holding. Typically, the farm site and the place of residence of the farmer coincide, but in case the farm site is very different from the residence of the farmer, address geocoding of the agricultural holding may produce erroneous results.

Another problem with agricultural holdings is that they represent the site of the holding rather than the spatial envelope of the farmstead. Area features, such as the agricultural land belonging to the holding, will be linked to a single point location. This may potentially cause erroneous outputs if, for example, administrative data on agricultural land linked to point locations is aggregated to grid cells, as a result of which the whole land will be assigned to the grid cell of the holding. The geocoding of agricultural holdings may therefore require alternative location strategies, using Land Parcel Identification System (LPIS)[27] or the collection of location data specifically for the purpose of geocoding the holdings.

In many countries, economic units or premises of industries or other enterprises can be geocoded by means of address information. In most cases, this will result in a decent spatial representation of the production site of the industry. However, the water discharge point of the industry may deviate significantly from the location of the production site, and as such cannot be properly represented by the address location. Using the address location may potentially cause erroneous outputs in the cases where water discharges should be aggregated from point-location to watersheds or river basin districts. Typically, locations of discharge points have to be retrieved from administrative registers on environmental permits or from monitoring systems.

In conclusion, all kinds of data have to be properly assessed with regard to the application purpose before geocoding. The possibility to assign data to a point location does not necessary mean that this option is advisable. The examples above demonstrate that using address locations to spatially represent an agricultural holding or industrial premises may be accurate in one context but not in another. This is the main reason why considerations regarding geospatial information must form an integral part of the design and production of statistics.

[27] European Court of Auditors 2016. Special Report No 25. The Land Parcel Identification System: a useful tool to determine the eligibility of agricultural land – but its management could be further improved.

6 Building and maintaining point-based geocoding infrastructure for statistical production

This chapter elaborates on the operational aspects of a point-based geocoding infrastructure. The aim is to provide guidance to support NSIs in the process of setting up and maintaining such an infrastructure. Guidance is targeting both NSIs which have not yet implemented this infrastructure as well as those who have it already in place but wish to improve the performance of their production.

The project has identified a number of key tasks to effectively set up, maintain and use a point-based geocoding infrastructure. The key tasks identified are:

- Find out what the users need
- Promote geospatial statistics and the potential of geospatial information
- Recognise geospatial data sources
- Assess data processing capacity
- Specify geospatial statistics output
- Create a flexible production set-up
- Build the geocoded survey frame
- Obtain and manage geospatial data
- Conduct geospatial data quality assessment
- Assess identifiers to enable correct data linkage
- Geocode data
- Prepare geospatial statistics products
- Assess data dissemination constraints

Ideally, these key tasks are to be implemented in a sequential order. However, in reality such ideal processes are rarely applicable. Some of the tasks may already be fully or partially implemented while others may not be relevant to all NSIs.

In essence, this chapter should be viewed as a cook-book containing a recipe for point-based geocoding, along with the proposed ingredients. Anyone committed to cooking knows that a cook-book can be used either as a strict step-by-step guidance to the final course or merely as a source of inspiration. The prospect of the project is that this chapter could be used as a step-by-step guidance, as well as an inspiring scheme to implement or improve a point-based infrastructure for statistics.

In order to demonstrate how the geocoding infrastructure can be set up and used as an integral part of the general statistical production process, each key task has been mapped against the main phases of the Generic Statistical Business Process Model (GSBPM), as shown in figure 8 below.[28]

The expectation is that this approach will help statisticians to better connect geospatial data management and geocoding to their professional reality. As the scope of the project has comprised the geocoding infrastructure itself rather than the use of geocoded data and the dissemination of geospatial statistics, Phase 6 (Analyse) is not fully pictured in this report, and Phases 7 (Disseminate) and 8 (Evaluate) are not covered at all.

[28] A more elaborate assessment of the GSBPM as a tool to mainstream geospatial information management in the statistical production is presented in Chapter 7.

Each key task is briefly described and, wherever possible and suitable, generic recommendations are provided. For each key task a selection of the relevant use cases have been identified and described, based on the experience gathered by the countries forming part of the project consortium. The use cases are presented in Annex 1.

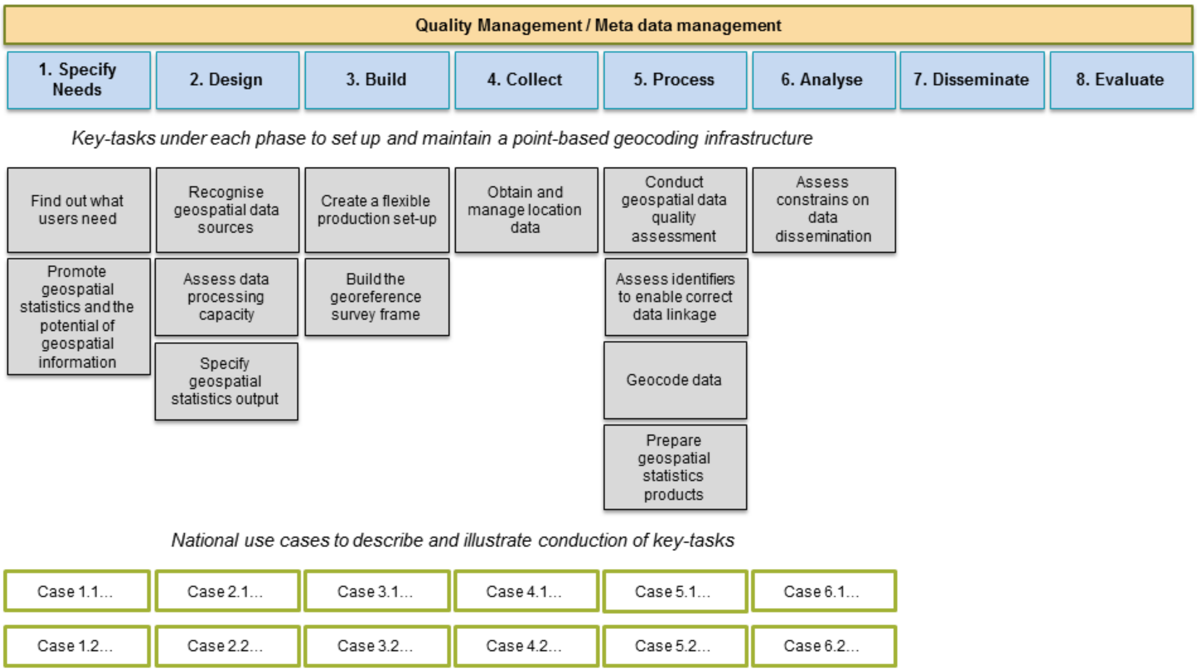
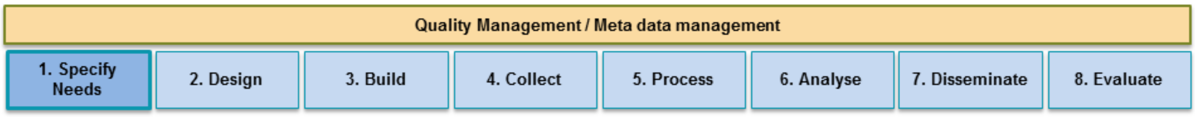

Quality Management / Meta data management
1. Specify Needs
2. Design
3. Build
4. Collect
5. Process
6. Analyse
7. Disseminate
8. Evaluate
Key-tasks under each phase to set up and maintain a point-based geocoding infrastructure
Find out what users need
Promote geospatial statistics and the potential of geospatial information
Recognise geospatial data sources
Assess data processing capacity
Specify geospatial statistics output
Create a flexible production set-up
Build the georeference survey frame
Obtain and manage location data
Conduct geospatial data quality assessment
Assess identifiers to enable correct data linkage
Geocode data
Prepare geospatial statistics products
Assess constrains on data dissemination
National use cases to describe and illustrate conduction of key-tasks
Case 1.1...
Case 1.2...
Case 2.1...
Case 2.2...
Case 3.1...
Case 3.2...
Case 4.1...
Case 4.2...
Case 5.1...
Case 5.2...
Case 6.1...
Case 6.2...

Figure 9: Key tasks mapped to the phases of the GSBPM.

6.1 Specify needs



The *Specify needs phase* is triggered when a need for new statistics is identified or the feedback on current statistics initiates a review. It covers all activities associated with engaging customers and users to identify their detailed statistical needs, proposing high level solution options and preparing business cases to meet these needs

A good understanding of users' needs is required in order to be aware not only of *what* to deliver, but also *when*, *how*, and, perhaps most importantly, *why*. Based on the user needs the geographical resolution of the output and as a result, the necessary spatial accuracy of the input can be determined, e.g. in the sample design. However it should be stressed that no matter what the actual output area is, all unit record information should be geocoded to the most accurate spatial resolution by default, as this facilitates the reuse in another context.

During this phase the integration of statistics and geospatial information can be promoted to users by explaining them the benefits. Consulting the stakeholders of geospatial statistics may be particularly challenging, not only because the users' community is heterogeneous and sometimes difficult to identify but also because the typical end-user of geospatial statistics consumes statistical information in a way that may be quite different from using traditional, non-geospatial statistics, involving GIS software and integrated spatial analyses. A population grid or census small statistical areas provides limited values unless it is analysed together with other geospatial information.

6.1.1 Find out what the users need

In order to make the right decisions on how to design the production setting, what thematic content to include in the geospatial statistics portfolio, what output areas to choose for dissemination, **it is strongly advisable to establish procedures for systematic consultations with the geospatial statistics users' community**. Consultations with the users can be conducted in numerous and more-or-less formalised ways. They can take the form of user councils, focus groups or information seminars, etc.

Use Case ***1.1 User councils in Statistics Sweden*** describes a formalised system for identifying users' needs through user councils, consisting of groups involving thematic experts and key users. The user councils meet twice a year. Use Cases ***1.2 Formalised user dialogue in Portugal*** and ***1.3 Formalised user dialogue in Austria*** describe similar mechanisms in Portugal and Austria where National Statistical Councils have been established in accordance with the statistical law, to provide formal room for dialogue between the producers and users of official statistics.

6.1.2 Promote geospatial statistics and the potential of geospatial information

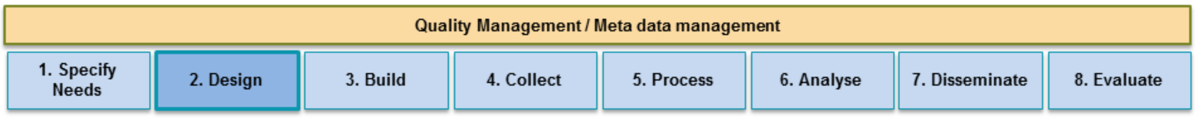
Identifying users' needs is an iterative process, involving both the users and producers. New, inexperienced or potential users cannot be expected to accurately articulate their needs for geospatial statistics without an active involvement from the producer.

The successful mainstreaming of geospatial statistics in spatial planning or other decision-making processes is typically the result of a decisive and active promotion of the potential of geospatial statistics by the producers. Hence, **it is of utmost importance to allocate adequate time and resources to promote geospatial statistics and to show-case its use, in order to identify new or emerging users' needs**. Participation in national or regional conferences to demonstrate the potential of geospatial statistics is usually a fruitful way to inspire a growing interest and awareness among various users' groups.

The promotion of geospatial statistics and geospatial information must not be limited to an external context. **Of equal importance is to raise the NSIs' awareness on how geospatial information can improve statistical production or add value to statistics by including the location context**.

Use Case ***1.4 Promote geospatial statistics and the potential of geospatial information*** describes Statistics Norway’s strategy to raise the awareness about the potential of geospatial data and statistics, both internally and externally, which includes participating in information seminars and lectures, as well as setting up an internal GIS resource centre.

6.2 Design



The *Design phase* typically describes the development and design activities, together with any associated practical research work needed to define the statistical outputs, concepts, methodologies, collection instruments and operational processes. When applied to a geocoding infrastructure, this phase usually occurs when the infrastructure is set up for the first time or needs to be improved.

High quality point location data lies at the heart of a point-based geocoding **infrastructure**. Accordingly, the choice of geospatial data on which the infrastructure could be built forms a crucial consideration to be made in this phase. An issue to be resolved is whether data could be obtained from an external provider or has to be collected by the NSI (see section 5.4). In the case of data collected by an external producer, agreements between NSI and the data producer may need to be concluded in order to safeguard the long-term provision of data. In addition, the ability to effectively manage geospatial data in terms of human resources and IT infrastructure needs to be addressed, along with the kind of geospatial statistics output to be produced.

It is recommended that NSIs should make a systematic assessment regarding the data sources, human resources and technical conditions needed to set up a point-based geocoding infrastructure. If such infrastructure already exists, the assessment may help to unveil needs for improvement.

6.2.1 Recognise geospatial data sources

The Global Statistical Geospatial Framework advocates the recognition of fundamental and authoritative geospatial data from the National Spatial Data Infrastructures or other nationally agreed upon sources.[29] According to the GEOSTAT 2 survey, geospatial data on address locations, buildings/dwellings and/or cadastral parcels form the complete basis for a point-based geocoding framework for statistics.[30]

Fundamental and authoritative geospatial data from the National Spatial Data Infrastructures is typically maintained under the authority or supervision of NMCAs. Local and regional administrations may also be involved in data collection, but in most cases the NMCAs gather and store data from municipalities in centralised repositories. In some countries, NSIs have established a direct collaboration with the municipalities providing location data to statistical offices.

Typically, geospatial data is of good quality if it is regularly used, e.g., for administrative purposes or business activities. Citizens usually have an incentive to provide correct and up-to-date addresses to administrations as they can expect benefits and services in return, such as health care, tax refunds or

[29] United Nations Expert Group on the Integration of Statistical and Geospatial Information 2016. Background Document on Proposal for a Global Statistical Geospatial Framework. Advanced Draft as of 28/07/2016.

[30] EFGS/GEOSTAT 2 2016. Spatialising Statistics in the ESS. Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs.

social benefits. Hence, address registers or other geocodes used for statistics should ideally be the same as for the administration, i.e. one single unique address register for all applications.

The following datasets are considered possible candidate sources for point-based geocoding:

- **Address data** – An address is a textual description of a physical point on the surface of the earth. Typically, it is either a postal address to which mail is delivered, or a physical address, which is the actual location of a person, household or business. The postal address location can be very different from the physical address location. For example, a property may have a post office box in the nearest town, which can be many kilometres away. In these and other similar instances, the postal address is not a good representation of where the actual property is located. Therefore, physical addresses are preferred in terms of identifying the location. A full street address is required for the accurate and efficient geocoding that produces a location coordinate (i.e. X and Y coordinate).[31] Physical address location data is commonly considered the most universal data source for point-based geocoding. There are several approaches to define the physical address location, e.g. the centroid of the building, at the entrance to the building, the centroid of the cadastral parcel. For different use cases it may be useful to maintain several physical address points that are spatially and conceptually related to each other and are based on building, dwelling and cadastral data (see below).

  Although not exactly defining specific points, postal codes, forming part of addresses, constitute important geocodes. Postal code areas or their centroids can provide very useful geocoding information, as for many stakeholders and data providers these are the geocodes which they are almost always aware of, contrary, e.g., to the local authority or NUTS code. One challenge related to this solution is that postal code areas can be the property of a postal services, i.e. often a private company. However, NSIs should have an unlimited and free access to such data, just like in the case of address data.

- **Building data –** According to INSPIRE, buildings are defined as constructions above and/or underground which are intended or used as a shelter for people, animals and things, or in the production of economic goods or the delivery of services. This term refers to any structure permanently constructed or erected on its site.[32] A building can be represented geographically as a polygon or a point. A point representation of a building is typically based on the central coordinate (centroid) of the building, or the location of the building entrance. For functional reasons, a real building can be divided into building parts, each part having its own geographical representation.
- **Dwelling data –** A dwelling is a room or a suite of rooms – including accessories, lobbies and corridors – in a permanent building or a structurally separated building part which, considering the way it has been built, rebuilt or converted is designed to be inhabited by one private household all year round. A dwelling can be either a single-family dwelling in a stand-alone building or detached edifice, or an apartment in a block of flats. Typically, a dwelling does not have a spatial representation of its own but is rather represented by the location of the building

[31] Australian Bureau of Statistics 2015. SSF Guidance Material – Geocoding Unit Record Data Using Address and Location.

[32] European Commission 2013. INSPIRE D2.8.III.2. Data Specification on Buildings – Guidelines.

to which it is associated. As such, dwelling information typically comprises an integrated part of building information systems or building registers.

- **Cadastral parcel data** – According to INSPIRE, a cadastral parcel is an area defined by cadastral registers or equivalent. A cadastral parcel should be considered a single area of Earth surface (land and/or water), national law under homogeneous property rights and unique ownership, property rights and ownership being defined by national law. Unique ownership means that the ownership title to the whole parcel is held by one or several joint owners.[33] A cadastral parcel can be represented geographically as a polygon or a point. A point representation of a cadastral parcel is typically based on the central coordinate (centroid) of the parcel. A cadastral parcel can contain one or more buildings and/or address locations.

In addition to the above-mentioned point-based data sources, **Road Network data** may constitute a valuable complementary source for geocoding. According to INSPIRE, a network is a collection of network elements, including mainly links and nodes, which create link sequences that represent a continuous path in the network without any branches. The road network links can be used to retrieve a location (e.g. an address location) in the cases where point-location data does not exist.

Depending on various traditions throughout Europe, the choice of a dataset for point-based geocoding may vary between countries. In some Member States, location data frameworks comprise integrated combinations of address information, building/dwelling data and cadastral parcels (see figure 10 below). Ideally, these objects are consistently and hierarchically linked to each other, both conceptually and topologically, which enable the inclusion of all three object types in the geocoding infrastructure. Yet, in other countries only building data or address data exists.

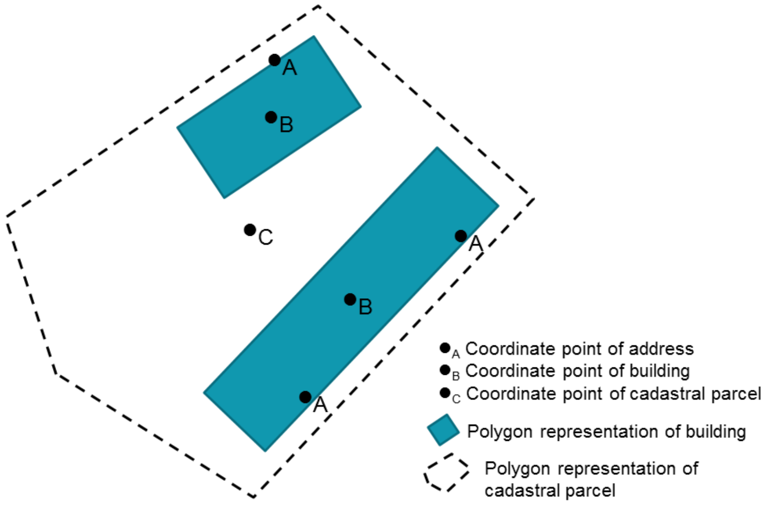

A
B
C
A
B
A
●A Coordinate point of address
●B Coordinate point of building
●C Coordinate point of cadastral parcel
Polygon representation of building
Polygon representation of cadastral parcel

Figure 10: A conceptual illustration of an integrated and hierarchical location data framework. Coordinates of buildings (B) and addresses (A) are linked to the cadastral parcel in which they are located. The cadastral parcel can be spatially represented by its centroid coordinate (C) or by a polygon feature delimitating its extent. The coordinate of an address (A) is linked to the building (B) to which it belongs (typically entrance). A dwelling does not have a spatial representation of its own, but can be linked to building and/or address location.

[33] European Commission 2009. INSPIRE D2.8.I.6. Data Specification on Cadastral Parcels – Guidelines.

A quite technical though important aspect is the choice of coordinate reference and projection system for the data in the geocoding infrastructure. If the geocoding infrastructure should not only support national use cases but also European use cases, the location data may need to be reprojected to a harmonised European coordinate reference system based on ETRS89 before aggregating data to European geographies. For European statistics, often the ETRS89-LAEA equal area projection is required as it allows the aggregation of point-based data into harmonised EU wide statistical grids such as the GEOSTAT 2011 population grid using the same projection. In the GEOSTAT 1B final report different approaches related to translation of data between national and European reference systems are discussed more in detail.[34]

The GEOSTAT 2 project advocates that a point-based geocoding infrastructure should be built on physical address and/or building data, but the choice of data must rely on the specific conditions in each Member State. Hence, the choice of data should be informed by the following generic recommendations:

- Completeness and coverage – Make a proper assessment of the completeness and geographical coverage of the datasets.
- Existence of sustainable maintenance policies – Make sure that the provider or custodian has a consistent and systematic scheme of maintenance to keep the dataset updated.
- Existence of data specifications and metadata – Make sure that the dataset is properly described in line with metadata standards (e.g. INSPIRE) or any other official data specifications.
- Consistent use of geocodes or identifiers – Make sure that the spatial objects in the dataset can be unambiguously identified by means of a nationally and officially agreed coding system.
- Suitability of location data – Make sure that the choice of geocoding objects is suitable for geocoded Census data may be accurately linked to specific locations using building or dwelling data but this does not necessarily apply to enterprises or industrial premises where data on physical address locations may be needed.
- Spatial accuracy – Make sure that the spatial representation of the location data is appropriate with the respect to the output area to be used for dissemination. Although the geocoding of statistical information does not typically require very high accuracy (the minimum accuracy of a couple of meters is enough in most cases), be aware that using the centroid coordinate of a vast cadastral parcel, or a very big building, may cause erroneous results if data is to be aggregated to a very fine grid (i.e. 100x 100 or 250x250 meter grids).
- Coordinate reference systems – Make sure that your geocoding infrastructure supports the European coordinate and projection systems based on the ETRS89 datum in addition to your national coordinate reference and projections systems.

All the above-mentioned aspects need to be holistically considered when assessing the usefulness and accuracy of a dataset. A dataset may be complete in terms of geographical coverage, but unless it complies with the official standards for the unambiguous identification of spatial objects, it may be useless for geocoding of statistics.

[34] EFGS/GEOSTAT 1 2012.

In some countries, access to geospatial data is regulated by the Statistical Act. Such acts may give access to all relevant administrative data, including geospatial data, free of charge or for a small fee. **If the Statistical Act does not regulate access to geospatial data, bilateral or multilateral agreements may be needed to ensure the regular and long-term access to geospatial information.**

Use Cases ***2.1 The Swedish National Geodata Cooperation*** and ***2.2 Data access through Norway Digital*** demonstrate the impact of national geospatial data agreements on statistical production in Sweden and Norway. Use Cases ***2.3 Legal basis for data and geospatial data access in Portugal*** and ***2.4 Legal basis for geospatial data access in Austria*** illustrate how geospatial data access is provided under the statistical laws in Portugal and Austria. The use cases also underline that formal agreements need to be supported by an active dialogue between NSIs and producers of geospatial information. Use Case ***2.5 Acquisition and processing of geospatial reference data in Poland*** describes an alternative strategy by Statistics Poland to set up their own address location database as the Polish NMCA did not possess a complete register for physical addresses.

6.2.2 Asses data processing capacity

In order to effectively manage geospatial information, a proper assessment of human resources and IT infrastructure should be conducted. Geospatial data management typically require a somewhat different expertise and technical infrastructure from what is mainly found in NSIs.

6.2.2.1 Human resources

A full integration of geospatial information into the statistical production process should be supported by bringing geospatial experts in close contacts with production teams. This may include organisational changes. There are different models to locate geospatial activities and GIS work within the organisation, and no clear success pattern can be identified. Either geospatial experts are directly involved in the production of statistics within the various sector-specific statistics teams (environmental, demography) or a crosscutting GIS unit or GIS centre of competence exists to support the production, often located in the IT or methodological and information systems department. The allocation of human resources for geospatial data management must be done with regard to the existing structure of the organisation. Accordingly, it is difficult to provide any generic guidance apart from stressing the fact that **high quality geospatial statistics require a permanent team of geospatial experts willing to support all steps of the production process,** including the close involvement and a working team established jointly with database administrators and server experts.

6.2.2.2 IT infrastructure

NSIs use various tools for the processing of geospatial data, building on standard GIS and database software. In most cases these tools, enhancing the features of standard products, have been developed over years and are specific to the production process in a given NSI. As for human resources, the diversity of technical environments in different NSIs prevents any generic recommendations to be made on specifications for the desired infrastructure and processing capacity. **However, it should be borne in mind that frequent production of geospatial statistics (e.g. on an annual basis) requires an industrial production setting**. It should also be recognised that the editing or processing of geospatial data may be different from dealing with traditional statistical or administrative micro-data, in terms of both methods and software. **NSIs need to prepare for investing in technical environments suitable for filling, processing or providing geospatial**

information that goes beyond standard equipment as regards computing resources, network performance, storage space, etc.

Use Case ***2.6 GIS Resource Centre in Statistics Norway*** reflects an ambitious attempt to raise the profile of geospatial information in statistical production by establishing a GIS resource centre in the Norwegian NSI. Use Case ***2.7 Resource setup in Statistics Finland*** describes how the implementation of a GIS strategy led to staff relocation. Use Cases ***2.8 Resource setup in Statistics Poland*** and ***2.9 Resource setup in Statistics Portugal*** describe the organisational and human resource backgrounds found in Statistics Poland and Statistics Portugal, with regard to the needs of geospatial information management.

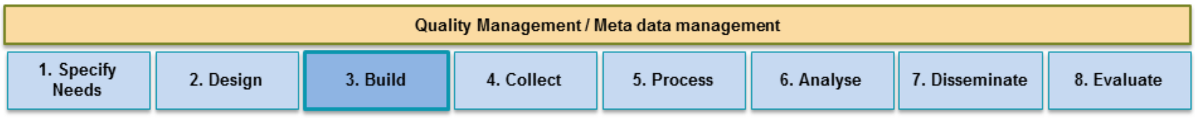
6.2.3 Specify geospatial statistics output

A fully-fledged implementation of a point-based geocoding infrastructure for statistics brings about vast opportunities for geospatial statistics outputs, in terms of thematic content and spatial outputs. In many national statistical systems, as well as in the ESS, geospatial statistics products do not yet belong to the portfolio of Official Statistics that is normally regulated by law. As a result, they may not be covered by the adequate appropriations and NSIs often have to recover, at least partly, the cost of their production by offering them to the market.

But even in those countries where production of geospatial statistics is mainly based on chargeable services, the strong push for Open Data has encouraged NSIs to offer some of their products free of charge.

It is recommended that **NSIs should consider a combined approach of mid-resolution products, including 1km^2 grids, based on Census indicators as a core set of Official Statistics or Open Data for the widest possible use,** in combination with the provision of chargeable services for high-resolution data or tailor-made services, as to increasing thematic content.

6.3 Build



In terms of geospatial data for geocoding of statistics, the *Build phase* is primarily about setting up and configuring a production environment, taking into consideration the long-term need for smooth maintenance and flexible output production. When applied to a geocoding infrastructure, this phase usually occurs when the infrastructure is set up for the first time or needs to be improved.

It is recommended that all NSIs should have a methodological description of their geocoding infrastructure. If this does not already exist, it should be the first follow-up action in the context of this report. A template on how to approach such a methodological description can be found in Use cases 3.1, 3.2 and 3.3.

6.3.1 Create a flexible production set-up

A flexible production environment is crucial for the efficiency and performance of the production process. A common approach among those NSIs that already have a point-based geocoding infrastructure in place is to set up a central “geodatabase”, “geography database” or “key code hub” storing references, at the level of point-locations, to any relevant statistical geography. Using such a model, an unlimited number of spatial references can be stored and kept separately from unit record data. The only reference that needs to be stored with unit record data is basically the unique identifier linking a unit record to its corresponding location in the “geodatabase”. This setup contributes to a smother spatial reference maintenance process and would also allow supporting European output geographies, in addition to national ones, in a much more flexible manner by storing e.g. the European grid cell references with the point-location objects in the “geodatabase”. Whenever the statistical or territorial geographies change, updates need to be applied only to the “geodatabase”, instead of updating multiple records of unit record data (See Figure 11 below).

No matter if references to statistical or territorial geographies are stored with unit record data or kept in a separate “geodatabase”, the preparation of a complete set of geocodes (administrative units, grids, urban zoning, tracts etc.) at the point-location level is a very efficient way to industrialise the production setting. This means that, once location data is linked to unit record identifiers, tabulations to retrieve statistics for various geographies can be conducted without even using GIS software or support from geospatial experts.

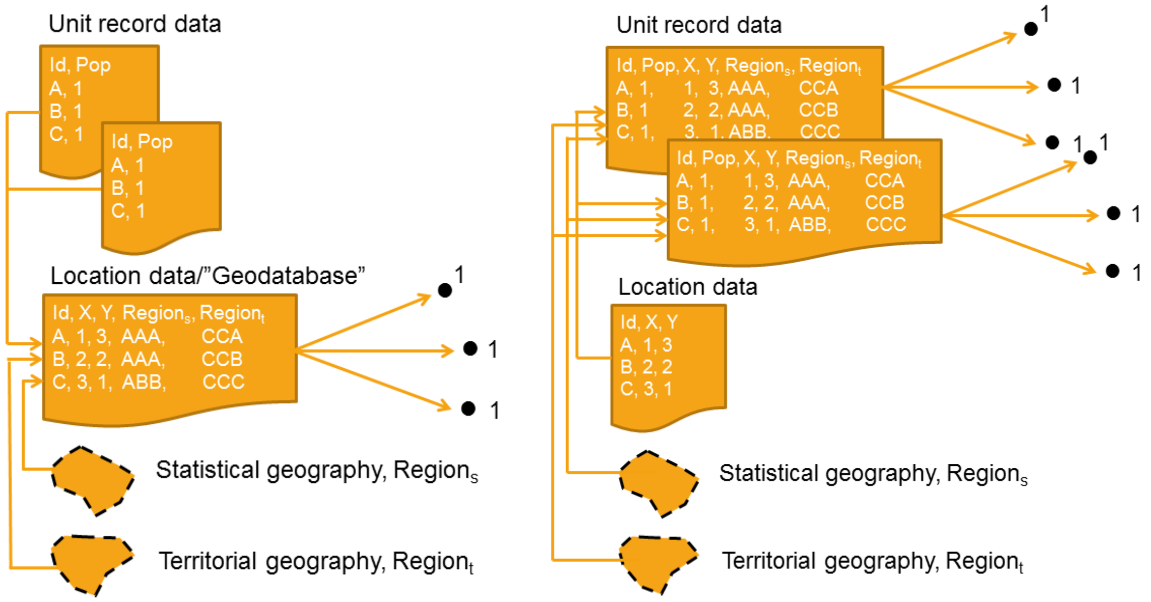


Figure 11: The model on the left-hand side shows the use of a "geodatabase" to store all references to statistical or territorial geographies. The only spatial reference needed in unit record data is the unique identifier, linking statistical information to specific locations. The model presented on the right-hand side illustrates the storing of references to statistical or territorial geographies with unit record data. Please note that Region$_s$ and Region$_t$ could also refer to different versions of the same territorial typology (time versioning).

Use Cases ***3.1 The Geography Database – the production set-up for point-based geocoding in Statistics Sweden***, ***3.2 The GeoDatabase – the production set-up for point-based geocoding in Statistics Austria*** and ***3.3 The production set-up for point-based geocoding in Statistics Portugal*** briefly describe and schematically illustrate an industrialised production setting in three different countries. The cases demonstrate both similarities and differences between the countries concerned.

6.3.2 Build the geocoded sampling frame

Recently many NSIs have started to geocode their sampling frames according to geographical nomenclature codes and by adding the exact or estimated spatial location of each record[35]. In the survey sampling context, the position of statistical units proves useful for three main reasons:

- It may help better organise the field work for face-to-face surveys.
- It may help spread the selected units over the population surveyed, which can be an efficient strategy for more accurate estimations.
- At last, at the cost of theoretical and practical issues, a geocoded sample may enable small area estimations or flexible output geographies other than the initial one.

Constructing Primary Units

Organising field work for the interviewers is one of the main issues related to face-to-face sureys. It entails the need to minimise travel costs, either in terms of time spent to reach each of the selected

[35] Benedetti, R *et al*. 2015.

units or in terms of money. A multi-stage sampling design appears to be a classic solution. As the first stage, geographical primary units (PU) are selected and assigned to an interviewer. Statistical units are then selected within each PU as the second stage. Each PU has to be as small as possible, while gathering enough statistical units to meet the interviewer work load. When the sampling frame is not geocoded, the PU consists in gathering some of the smallest geographical units available. The latter may not be suitable. With a geocoded sampling frame, the PUs can be constructed automatically at a lower cost (see Use case 3.4 from France).

Improving the sampling design

In the literature regarding survey methodologies, one of the primary topics of interest is how to improve estimations of the population characteristics using some additional knowledge of the sampling units. The gains are even noticeable when the enhancements are applied to the sample design, rather than to the estimator.

Using the location of statistical units in the sampling design has attracted a lot of interest in recent years. There are certain theoretical criteria, and their empirical counterparts, to assess whether a spatial sampling design will prove useful, for more information see Use Case 3.4.

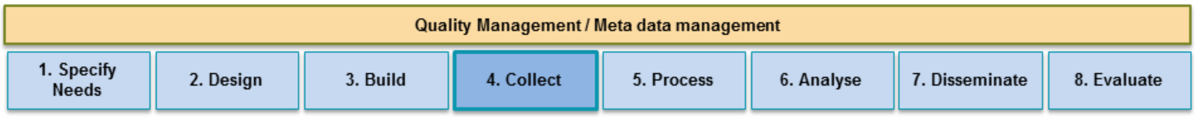
Disseminating small area statistics

From a theoretical point of view, a geocoded sampling frame and, in consequence, a geocoded sample, makes it possible to disseminate results on any small area, which is more generally referred to as a domain. From a practical point of view, the accuracy of such estimations might be very poor.

In order to assess a variety of new statistical methods available with a point-based geocoding infrastructure, INSEE is currently, with grants from Eurostat, developing a handbook of spatial statistics to be released by the end of 2017 or at beginning of 2018. The issue of spatial sampling fully falls within the scope of the handbook, and as such will be dealt with more precisely.

Use Case ***3.4 The Labour Force Survey (LFS) sample in INSEE*** describes how the sampling frame for the Labour Force Survey in France is created to ensure a spatially well-distributed sample. Use Case ***3.5 Sampling Frame – National Dwellings Register*** presents a similar approach to creating a sampling frame for social surveys in Portugal. Use Case ***3.6 The sampling frame for social surveys (SFFS) in Poland*** describes how the address and dwelling sampling frame, created for the purposes of the 2011 National Census, is used to enhance sampling for social surveys (SFFS) in Poland.

6.4 Collect



In the *Collect phase* all necessary information (data and metadata), are gathered using different collection modes, and loaded into the appropriate environment (created in the *Build phase*) for further processing. When applied to a geocoding infrastructure, in contrast to the previous phases, this phase occur each time new geospatial data is obtained or collected.

Location data is typically acquired from external producers, rather than being collected by the NSI in the sense of direct data collection. However, in some Member States the collection of location data, such as buildings and address locations, is conducted under the authority of the NSI or as a joint venture involving the NSI, local government bodies and the NMCA (See the description of the "In-house" approach in section 5.4).

6.4.1 Obtain and manage location data

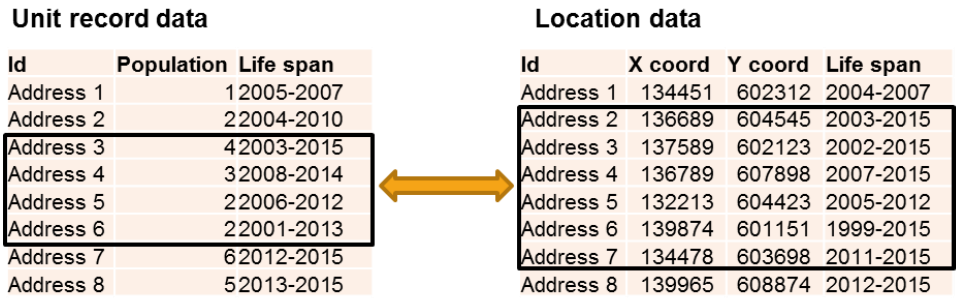
Procedures for the acquisition of location data by NSIs from NMCAs or other data providers may vary in terms of the frequency and mode of data transfer. Some NSIs obtain data manually, receiving complete copies of source data files for local storage, either annually or at other frequencies. Other NSIs have launched web services allowing data to be obtained by means of automatic notification procedures from the producer at monthly, weekly or daily intervals, or whenever an object is changed in source data. The launching of such services in Member States have been triggered by the implementation of INSPIRE.[36] The availability and easy access of metadata services allow Member States to recognise existing data. Even though data may not be harmonised according to the INSPIRE directive, the main tendency an increasing number of INSPIRE compliant services.

Regardless of the data acquisition mode, the key challenge for any NSI is how to deal with the temporal aspects of data in order to obtain the best possible temporal cohesion between location data, on the one hand, and statistical and administrative data to be geocoded, on the other. This is perhaps the most typical of those NSIs that use continuously updated administrative data sources (e.g. population registration) instead of fixed collections of Census data.

To manage the issue of temporality, it is important to work with date fields (the validity start-date and end-date) and status fields, in order to describe the status during the validity phase of a record. **Historic records should be kept, which means that records in registers should never be deleted, but their status should be set as 'end' and the end-date should be given.**

A common approach is to create situational extracts of the location dataset and the unit record data (administrative data), both corresponding to the desired reference time. To retrieve such situational extracts or frames from living records or databases, a consistent use of time stamps is required. **If the life-span of each location data object is consistently declared, the database can be rolled-back to create situational extracts for any desired point of time**.

[36] The INSPIRE Thematic Cluster for Addresses may provide a useful input regarding setup of services in Member States. https://themes.jrc.ec.europa.eu/groups/profile/1849/addresses



Unit record data

Id	Population	Life span
Address 1	1	2005-2007
Address 2	2	2004-2010
Address 3	4	2003-2015
Address 4	3	2008-2014
Address 5	2	2006-2012
Address 6	2	2001-2013
Address 7	6	2012-2015
Address 8	5	2013-2015

Location data

Id	X coord	Y coord	Life span
Address 1	134451	602312	2004-2007
Address 2	136689	604545	2003-2015
Address 3	137589	602123	2002-2015
Address 4	136789	607898	2007-2015
Address 5	132213	604423	2005-2012
Address 6	139874	601151	1999-2015
Address 7	134478	603698	2011-2015
Address 8	139965	608874	2012-2015

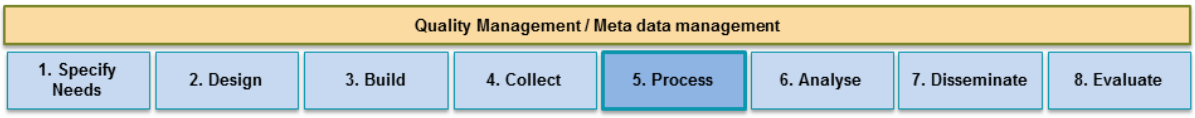
Figure 12: The figure provides a very simplified illustration on how to establish frames, or "situation extracts", representing the year 2011 for unit record data (population data), and the corresponding location data (address data) using the life-span information. Addresses 1, 2, 7 and 8 are excluded from the unit record data frame as no people were present at these addresses in 2011. In location data, only Addresses 1 and 8 are excluded from the frame as they are non-valid objects in terms of the reference time, according to the time-stamp. When merging unit record data with location data, Addresses 2 and 7 will return zero population as there is no temporal correspondence with the population in unit record data for the year 2011.

Location data used in the direct data collection (e.g. Census operations) should be ideally already available at the time of the creation of statistical unit record data objects. The data entry procedure should be designed for all registers so that the addresses could be compared with an authoritative index of valid addresses during the acquisition process. During the data entry, the right spelling should be checked and, ideally, only the valid addresses should be accepted, so that a direct match between the address location dataset (with coordinates) and the unit record data could be established. The address dataset must always be as timely as the case processing requires. The address of a new building, for example, including a correct notation of the street name and coordinates, associated with the admission of the first resident in the register the corresponding address, must be available.[37] In some cases, permits for new buildings can have a provisional address for a shorter or longer period, even after the residency permit is issued by the legal authority. In these cases, address updates must be guaranteed in the appropriate time, in order to avoid duplicate geocoding of the same building or mismatches between location data objects and statistical unit records.

Use Case ***4.1 Managing temporality in the Building register*** provides some generic guidance on how the temporal aspects can be handled, based on Statistics Austria's maintenance of its Building register which is a living record that changes from day to day. Use Case ***4.2 Address point acquisition in Statistics Poland*** describes the work conducted by Statistics Poland to create its own location dataset as no complete set of address location data existed at the time of the 2011 Census round.

[37] Haussman, M et al 2016.

6.5 Process



The *Process phase* describes data cleaning and preparation for analyses. It consists of several sub-processes that check, clean, and transform the input data, so that it could be analysed and disseminated as statistical outputs. From the geospatial information perspective, this is by far the most laborious phase. This is the phase where the main quality assessments are conducted on the geospatial data obtained in the *Collect phase*. It also comprises the integration of geospatial data sources and statistical or administrative data, e.g. through geocoding.

6.5.1 Conduct geospatial data quality assessment

Just like any information obtained for statistical purposes, geospatial information needs to undergo various procedures to secure the quality of the dataset. The scope and complexity of such procedures may vary depending on the origin of the data and the degree to which it has already been sufficiently assessed and checked by the custodian or data provider. Due to the spatial element, assessing the quality of geospatial data may be conceptually different from the assessment routines traditionally applied to statistical information although semantic and structural quality assessments of the attribute information are often similar. Typically, methods and procedures for geospatial data assessment are poorly described in guidelines for statistical production.

In a basic sense, geospatial data quality assessment is about finding out if the location of a spatial object (defined by its X and Y coordinates) is true and accurate. There is a range of more or less advanced methods to conduct such assessments. Selected basic approaches are described below:

- Crosscheck point-datasets with a polygon dataset accurately delimitating the territory of the country. Any point that falls outside the polygon is invalid.
- If the point-objects have been assigned to territorial units (by means of a territorial code) the point-dataset may be crosschecked with a polygon dataset on territorial units (including postal code areas). If the original territorial code is different from the code assigned through the point-in-polygon operation, there is a strong probability for erroneous coordinates.
- Point-datasets can also be crosschecked against high-quality polygon data on land and water. If a physical address or a building point falls into water there is a strong probability for erroneous coordinates.
- If data is provided in time series, procedures for comparing against last year's data can be conducted. If there are big differences (in location), potentially erroneous objects can be rejected and presented in a table for further inspection.
- Topological consistency might need to be checked, e.g. an address point on the correct side of the street or other topographic features (rivers, etc.). It is important if the street separates enumeration areas.

As a rule of thumb, it is better to establish routines for data verification when data is being put in to the record or database than trying to fix the errors afterwards. Data editing by the source is preferred before *ad hoc* correction of data. The obvious reason for this is that *ad hoc* correction produces manipulated versions of source data that need to be maintained as parallel systems.

However, as location data is typically collected by NMCAs and/or municipalities, NSIs may have limited possibilities to correct errors by the source. **To overcome this problem it is recommended to establish systematic feedback routines between NSIs and data providers for the reporting of errors to be corrected by the custodians.**

Use Cases ***5.1 Cross-check and geospatial data validation*** and ***5.2 Geospatial data assessment*** describes different procedures for the validation of geospatial data conducted in Statistics Portugal and Statistics Norway.

6.5.2 Assess identifiers to enable correct data linkage

One of the key features of a point-based geocoding infrastructure is a consistent use of identifiers. In order to safeguard correct linkage between location data and unit record data to be geocoded, unique and consistent codes able to unambiguously identify spatial objects (physical address location, building, dwelling or cadastral parcel) are of paramount importance.

At this stage of the process (before identifiers in location data can be compared with unit record data), consistency of identifiers can only be checked by means of logical rules with regard to the coding system itself. Hence, it is not possible to assess whether the identifiers will match the identifiers of the unit record data to be geocoded at this stage.

Relevant procedures can be as follows:

- Looking for and correcting duplicates.
- Checking the consistency of address strings in accordance with the official data specification (street name, number, etc.).
- Checking the consistency of any other identifier (checking the number of positions and the expected composition of the code).

Another important step is to assess the life-span of the identifier (e.g. an address or a building ID) to declare if it is active or not. Such a procedure may differ from country to country depending on data models and procedures for maintenance of source data on physical address locations or buildings, etc. In a living record, an object can be either discarded if it is no longer active, or it can be kept in the record but flagged with a time-stamp indicating that its life span has expired, e.g. ceased to exist.

In essence, it is important to recognise how the temporality of data is managed by the data custodian in order to know how to build a consistent point-based geocoding infrastructure.

As for the assessment of geospatial data quality, it is strongly recommended to establish feedback routines to report to custodians errors found when assessing identifiers.

Use Cases ***5.3 Non-spatial data assessment*** and ***5.4 Maintaining the quality of a point-based building register*** illustrates how the quality of identifiers in location data is assessed and maintained in Statistics Norway and Statistics Austria. Use Case ***5.5 Quality indicators for non-spatial data*** describes a system of indicators used by Statistics Portugal to assess and describe the quality of identifiers to link data.

6.5.3 Geocode data

Geocoding is the process of assigning a geocode to a piece of information (e.g. a statistical unit record) using known location information, such as coordinates. If unit record data and geospatial location data have been prepared, in accordance with the tasks described in previous phases, with consistent and unambiguous identifiers, linking statistical information with point location is a relatively straightforward process. Data pairing can be conducted in many different ways, on various technical platforms and with various outputs. Unit record data can be paired with location data into virtual tables where aggregations to various statistical geographies are performed simultaneously in the same step. The result from geocoding can also be stored in physical point-data layers for further processing and spatial analysis.

Geocoding can be conducted by simply joining location data with unit record data within a database environment (such as SQL, PostgreSQL, Oracle, etc.) which is part of regular IT infrastructure in many NSIs. It can also be conducted in a desktop GIS environment, as most standard desktop GIS software have built-in geocoding capabilities for linking table data to address locations. Some NSIs have implemented use of web services and APIs with location data, set up by NMCAs or other providers, in their geocoding practises, while other use locally stored tables or spatial databases. Automatic geocoding is typically split into three separate components:

- Parsing – extracts relevant parts of the reported address into separate fields to allow matching.
- Matching – matches the parsed address information to a coding index.
- Coding – allocates a geocode and match information to each reported address.

This is an iterative process; the addresses may pass through these stages a number of times, until the highest quality match and geocode is found. With each pass the address information is parsed differently based on the range of possible matches identified in previous passes.[38]

The true challenge of high quality geocoding comes with inconsistent or erroneous identifiers resulting in mismatch between location data and unit record data. **It is recommended that NSIs develop strategies to deal with these problems in a consistent and harmonised way.** For any data collected and maintained under the authority of the NSI, correction of the data by the source is the most preferred solution (e.g. correction of non-valid addresses reported by companies or individuals). But if mismatches occur due to errors in location data or administrative data sources obtained from other custodians, the approach to edit or correct data by the source is more difficult or may not be applicable at all.

Alternative strategies to consider are different types of *ad hoc* correction including address validation tools, homogenisation of address information or interpolation of address location points. A common approach is a step-by-step matching procedure sequentially pinpointing the unit record data to location, progressively less precise in spatial terms (from full address match to match only on street name to zip-code, locality or district, etc.). Such procedures may include the use of non-point based data such as postal code areas or road networks.

[38] Australian Bureau of Statistics2015. SSF Guidance Material – Geocoding Unit Record Data Using Address and Location. Second release.

It is strongly recommended that metadata be provided on the quality of geocoding at individual record level. From this metadata it should be possible to interpret the accuracy at which the information has been geocoded.

In Use Cases ***5.6 Geocoding population data in Statistics Sweden***, ***5.7 Geocoding workplaces in Statistics Sweden***, ***5.8 Geocoding workplaces in Statistics Portugal***, ***5.9 Geocoding practise in Statistics Finland*** and ***5.10 Geocoding practise in Statistics Austria*** a number of examples are presented on the approach used for the geocoding of data in different Member States. Problems occurring due to mismatch between location data and unit record data are described as well as generic solutions.

6.5.4 Prepare geospatial statistics products

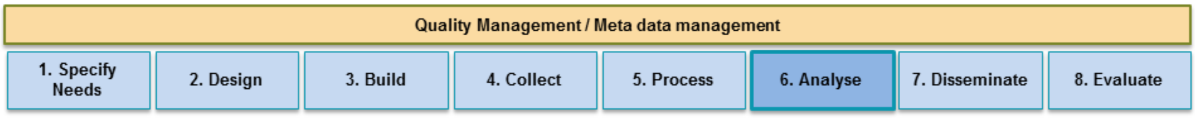
This is the phase where the production of the final product according to the user needs takes place and the user can assess if the product meets the expectations (see section 6.1.). A point-based geocoding infrastructure brings about a great spectrum of output opportunities and as a result spatial adjustments of the final product to user requirements and change request are fairly straightforward when compared to a traditional non geocoded production design. Non-aggregated geocoded unit record data can be used to enhance spatial analyses internally, but most notably, a variety of geospatial statistics products can be retrieved by means of aggregations to different statistical geographies. Depending on the production setup, preparation of geospatial statistics products starts already in the geocoding process, as data can be geocoded and aggregated simultaneously in the same step.

Other aggregations may require GIS based operations where geocoded unit record data is processed together with polygon-based data (buffering, intersection, etc.). The preparatory stage of grid statistics by means of different aggregation methods is well documented in the guidelines produced by the GEOSTAT 1 project [39] and can be found at the www.EFGS.info website. Many elements in these guidelines are generically relevant for the production of any regional or small area aggregations.

Use Cases ***5.11 Geospatial statistics portfolio in Statistics Poland*** and ***5.12 Geospatial statistics portfolio in Statistics Austria*** describe the different geospatial statistics products and services provided in Poland and Austria. Use Case ***5.13 Metadata and INSPIRE compliance*** describes the services made available by Statistics Finland along with a description on how INSPIRE compliant Metadata is provided.

[39] EFGS/GEOSTAT 1 2013. Production procedures for a harmonised European Population Grid. Aggregation method.

6.6 Analyse



In the *Analyse phase*, statistical outputs are produced, examined in detail and made ready for dissemination. It includes preparing statistical content (including commentary, technical notes, etc.), and ensuring outputs are "fit for purpose" prior to dissemination to customers.

The final production and use of geospatial statistics is not within scope of the GEOSTAT 2 project, accordingly the *Analyse phase* will not be exhaustively pictured in this report. However, a crucial task of the *Analyse phase* is to assess constraints on data dissemination, including confidentiality issues or legal issues related to use of geospatial data for production of statistical content. This aspect will be discussed in more detail as part of the GEOSTAT 3 project.

6.6.1 Assess constrains on data dissemination

In the presence of geospatial data, disclosure control experts must face a paradox. On the one hand, such data need more protection because they allow more identification, and on the other hand they offer many possibilities for analysis, that users don't want to distort too much by suppressing data.

Disclosure risk is higher when considering geospatial data:

- firstly, because belonging to a geographical area may give information to the intruder about some attributes (e.g. 100 percent of inhabitants of a square are unemployed). This is called categorisation risk, and it increases in the case of spatial data because of Tobler's "first law of geography" which states that "everything interacts with everything, but two close objects are more likely to do so than two distant objects";
- secondly, because of so-called identification risk. Indeed, among the characteristics shared with someone, a common geographic area leads to a higher probability of identifying the person (one probably knows better our neighbour than someone who one shares any other characteristic with). Moreover, identification of addresses has recently become possible with the development of open access tools like Google Street View. As a result, population density is a fundamental predictor of disclosure risk: the lower the density, the higher the disclosure risk. That is why confidentiality thresholds can differ between countries;
- finally, disclosure risk can increase with the geographic differencing issue, when data is disseminated at different levels (hierarchical or not).

Technically, the dissemination classification (zoning, administrative boundaries, or regular tessellations such as grid squares) is a categorical variable like any another one (an additional dimension of tabular data). It is therefore possible to deal with disclosure risk with no geographical consideration. Nevertheless, a geographically intelligent management of disclosure issues will preserve the underlying spatial phenomenon. A risk-utility compromise has to be made, using relevant distortion indicators.

A partial solution is to consider several types of users (usually researchers VS general users), and to disseminate as many files as types of users, with different levels of disclosure risks and different levels of perturbation associated.

Several NSIs or researchers have been working on the implementation of geographically intelligent disclosure control methods. They have identified two main strategies: pre-tabular methods, applied to the micro data (noise adding, swapping…) and post-tabular methods, applied to aggregated data (rounding, cells suppression...).

A review of the existing literature will be drawn up in a devoted chapter of the forthcoming handbook of spatial statistics, to be published by Insee, with the support of Eurostat, by the end of 2017 or beginning of 2018. This chapter will also contain different practical use cases, one largely inspired by work for the Eurostat Grant « Harmonized protection of Census Data in the ESS ».

Very practically, dealing with spatial micro-data adds a layer of complexity in the disclosure control process because it involves very large data. More details on how to deal with large arrays of micro data when protecting privacy can be found in the use cases.

Use Cases ***6.1 Data dissemination in Portugal***, ***6.2 Dissemination of grid data for tax information in France: issues and solutions*** and ***6.3 Constraints on data dissemination in Austria*** all address considerations to be made with regard to disclosure and other constrains related to the dissemination of data.

7 The GSBPM as a tool to organise, set up and use a point-based geocoding infrastructure in statistical production

The aim of this section is threefold:

- To demonstrate how the GSBPM can help NSIs to mainstream the use of geospatial information in their statistical production process
- To offer suggestions on how the GSBPM could be extended to recognise better the use of geospatial information in the statistical production process
- To demonstrate how the GSBPM can be used to model the production of geospatial statistics.

One task of the GEOSTAT 2 project was to test and evaluate the GSBPM when geospatial data is involved in the statistical production process. The questions were: 1) is the GSBPM a usable tool to understand and describe the geospatial dimension of the statistical production process? 2) If yes, does the GSBPM need further development or does its documentation need more accurate expressions when geospatial data is possibly included in the process? The project aimed to achieve common views of subjects that could be handed to UNECE for evaluation who are responsible for maintaining the GSBPM.

The starting point for testing was that the GSBPM is implemented in many NSIs but typically not applied to describe and structure processes implying the use of geospatial information. As there are no references to geospatial data in the GSBPM's current guidelines, there was no given baseline to test the model in a comparable way. Lack of experience among the project members concerning methods on how to use the model made the task even more difficult.

The evaluation of the GSBPM has been conducted through a set of workshops and exercises conducted by the project consortium jointly and by the project members and co-staff in their respective NSIs. This section comprises the common conclusions and main findings from the national exercises. The full documentation from each country is presented in Annex 2.

7.1 The GSBPM as a tool to help NSIs to mainstream the use of geospatial information in their statistical production process – main benefits

Despite the fact that the current version of the GSBPM does not recognise the scope of geospatial information in the statistical production process, it provides a link to internationally agreed statistical processes and facilitates the communication between the statistical and geospatial communities. As such, the overall conclusion is that the GSBPM can play an important role in mainstreaming the use of geospatial information into the generic statistical production processes. It gives common frames and promotes common concepts and unified methods with other in-house processes, as well as equivalent processes in other organisations.

Benefit 1: Geospatial elements visible

One of the main benefits of the model is that it helps to see *where* and *how* geospatial information might appear, or should appear, in various phases of the production process. Geospatial information can be involved along the whole process or it may be only a part of one particular phase of the production.

Benefit 2: The role in the statistical production process visible

Secondly, it may also help to clarify that the role of geospatial information varies depending on the statistical domain. Geospatial information can be source data and the starting point for production, but it can as well be the outcome and final product of the process. When it is used to enrich statistical data by location information or by other geospatial elements, geospatial information plays the role of auxiliary data. In any case, when location is an important part of the statistical production process, it induces special demands and challenges to be solved.

Geospatial data and production phases bring additional dimensions to statistical production process. Figure 13 includes examples by the GSBPM phases of what these dimensions can contain. The illustration separates work phases focusing on plain statistics (bottom) from the work phases focusing on geospatial data (top). In-between lays the level of integration. The integration of statistics and geospatial data (as described in Section 4.1) requires established arrangements between stakeholders, adequate technologies, as well as standards in order to function through the process.

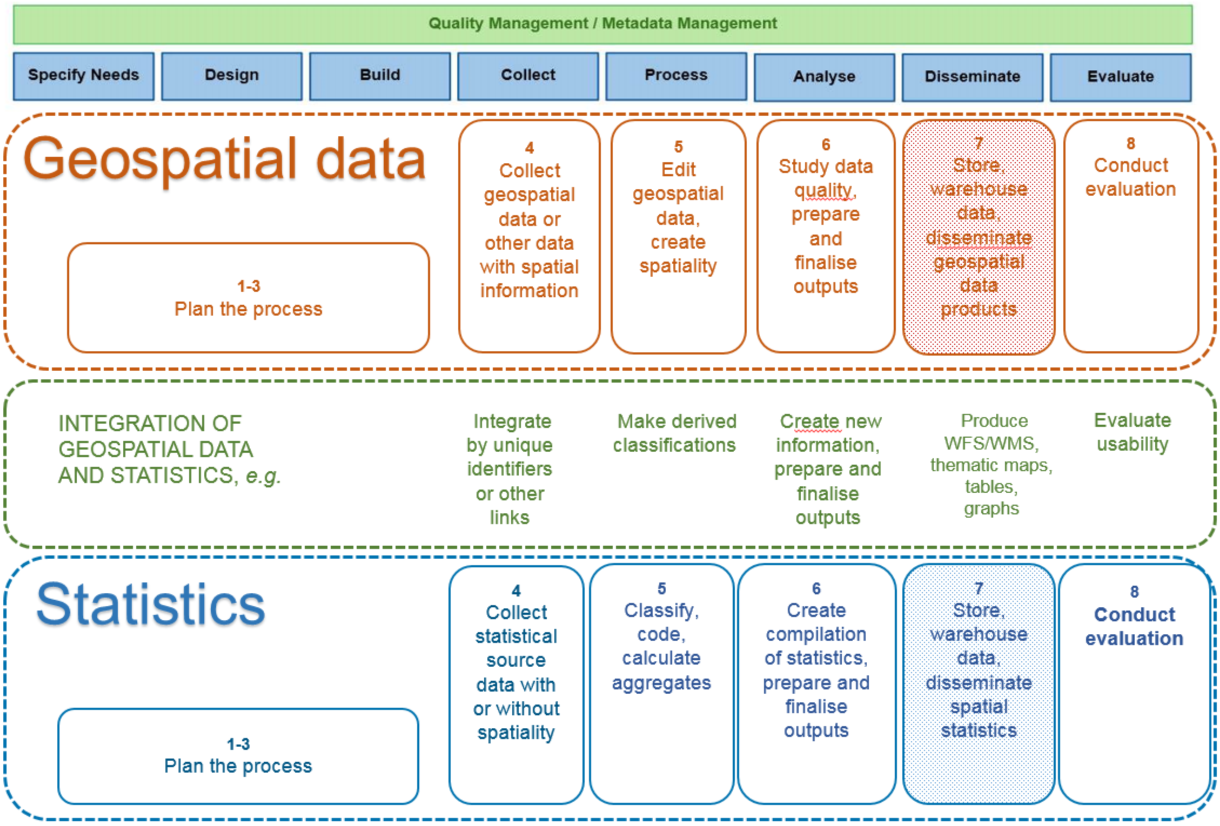

Quality Management / Metadata Management
Specify Needs
Design
Build
Collect
Process
Analyse
Disseminate
Evaluate
Geospatial data
1-3 Plan the process
4 Collect geospatial data or other data with spatial information
5 Edit geospatial data, create spatiality
6 Study data quality, prepare and finalise outputs
7 Store, warehouse data, disseminate geospatial data products
8 Conduct evaluation
INTEGRATION OF GEOSPATIAL DATA AND STATISTICS, e.g.
Integrate by unique identifiers or other links
Make derived classifications
Create new information, prepare and finalise outputs
Produce WFS/WMS, thematic maps, tables, graphs
Evaluate usability
Statistics
1-3 Plan the process
4 Collect statistical source data with or without spatiality
5 Classify, code, calculate aggregates
6 Create compilation of statistics, prepare and finalise outputs
7 Store, warehouse data, disseminate spatial statistics
8 Conduct evaluation

Figure 13: Dimensions in production of geospatial data

Benefit 3: Requirements visible

Thirdly, it may help to clarify that a production process where geospatial data is involved may require special skills and technical environments at different stages of the process. Both statistical and geospatial experts are usually needed. The input of different specialists at different process phases is illustrated in Figure 14 below. Certain phases may need more intense geospatial expertise than others. A geospatial expert is visualised in the figure by a person with green background. In Process 1, geospatial expertise is dominant. In Processes 2 and 3, the geospatial issues are parts of one or two particular process phases. The spatial related work phases may appear only once but the impact of that phase may be crucial for the success of the process from the point of view of bringing spatiality to the production of statistics. It should be noted that if processes are modernised in a systematic way, automated or service-based production may reduce the need for different types of professionals.

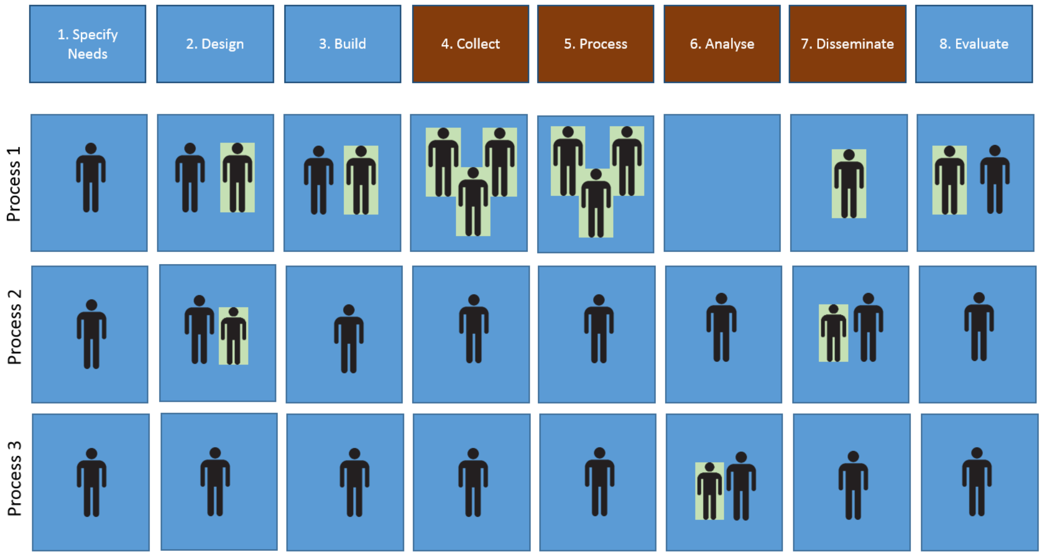


Figure 14: In the geospatial statistical production process, the need for geospatial data and expertise varies. Certain phases may need more intense expertise than others. A geospatial expert is visualised by a person with green background. In Process 1, geospatial expertise is dominant. In Process 3, the geospatial part of the process is a part of one particular process phase.

Benefit 4: Common language and understanding bring quality

The GSBPM, as a common framework, offers the possibility to describe commonly the production process. It may increase mutual understanding between the parties involved, providing a uniform way to communicate. Best practices, instructions and tools can be better shared. From an organisational point of view, the GSBPM offers a tool whereby separate processes are also able to communicate with each other. The geospatial dimension may be carried out in the same way or with a common tool in different processes. This is not only excellent in terms of efficiency, but also in terms of quality management. It is a quite common situation that stability management between processes is lacking.

7.2 Suggestions to improve the GSBPM to recognise better the use of geospatial information in the statistical production process – main findings

The GEOSTAT 2 consortium agrees on the fact that the GSBPM can play an important role to mainstream the use of geospatial information into the statistical production processes. Nevertheless, the current version of the model fails to support a number of vital tasks associated with geospatial data management in statistical production. This prevents a coherent use of the model.

The extension of the GSBPM model to recognise the use of geospatial information in the statistical production process would provide a much needed link to internationally agreed statistical processes and facilitate the communication between the statistical and geospatial communities. In addition, including geospatial information management in the GSBPM would extend the infrastructure for geospatial statistics to the actual data production and thus bring it to the core business process of statistical offices. The project advocates that UNECE take this work further under the High Level Group on Modernisation of Official Statistics.

The project members of the project consortium have arrived at slightly different conclusions (because of the different set up environments and data) regarding possible revision of the GSBPM. The following common conclusions can be drawn regarding the use and further development of the GSBPM:

1. When the GSBPM is actively applied to modernise a statistical production process, geospatial data and processes should be included.

1.1. If the GSBPM is used for the whole statistical process, both statistical and geospatial production phases should be described.

1.2. The GSBPM facilitates communication between different specialists who participate in planning and conducting a process. The model has the potential to be a common framework to understand the mutual roles of statistical and geospatial data and their integration. This would, however, require a stronger recognition of geospatial terminology in the model (also see 2.3).

1.3. The model helps to recognise phases which are common and can be processed with common tools and/or methods, or which can be, e.g. automated in order to increase efficiency.

1.4. The model helps to identify what is needed for geospatial statistical production, whom it may concern, what the preconditions are, and what the most critical phases to produce geospatial statistics are.

2. The documentation of the GSBPM is not adequate to describe a geospatially extended statistical production processes. The current guidelines need too much interpretation and they cannot be used as such. Supplementing of the documentation increases coherent and uniform use of the model.

2.1. Revision of the descriptive texts is necessary in order to cover geospatial data management. The proposals put forth in this report together with the national experiences gained through the national exercises in the GEOSTAT 2 could be used as a basis for changes and improvements. Detailed suggestions for text amendments are collected in Annex 4.

2.2 If geospatial activity cannot be embedded by revision of the text, then adding a new sub-process should be considered. (Adding a new sub-process does not overrule the need for text amendments in other sub-processes).

2. 3. The terminology used in the guidelines should be uniform and generally understood. The geospatial statistical community should be consulted when choosing the geospatial terms.

3. Use of other UNECE models for statistics may also be beneficial. The advanced models of UNECE should cover the geospatial dimension in order to cover the production of statistics as a whole.

3.1. A better understanding of production, managing and utilising the geospatial dimension for statistics may still require further examination and development of the models (the GAMSO, the GSIM and the CSPA)

3.2. Extensive support for implementing the models is needed both in-house and from UNECE.

7.3 Main findings of the national exercises

The project members started the work by getting familiar with the model. Lack of a common baseline for testing the model in a comparable way made the testing difficult. The equivalent use of the model would have required a mutually agreed way of using it in order to avoid different interpretations. However, the agreement would have required better understanding of the model and its use in specific situations. This is why the organisation's maturity to use and to apply the GSBPM took on a bigger role than expected. The amount and quality of in-house support for applying the GSBPM to geospatial related processes varied much based on experiences in testing, ranging from good support to very little support at all:

1) The GSBPM and geospatial specialists were found, the discussions were fruitful, their support was on the right level to promote the work
2) The right persons were found, discussions were partly fruitful but practical results were missing. Applying the GSBPM seemed to be too big a workload
3) The GSBPM experts were found but discussions produced no proper benefits, no support for applying the GSBPM
4) The project member could not find any support for the work

Tests were conducted in Finland, Sweden, Austria, Poland, Portugal and France. Despite the fact that the project did not achieve a completely similar level of approaches, it was possible to find common or comparable views of the GSBPM's phases (but possibly not at the level of sub-processes). Even though there are the slightly different approaches, every view on the GSBPM reflects the reality of the national production environment and data sets and they are all extremely relevant and therefore should all be taken into account.

7.3.1 Summary of findings by NSIs

Below is a short summary for each country. More precise results of the national exercises are collected in Annex 2.

The production environment and NSIs' role in managing geospatial statistical data differs from country to country and causes large differences in the processes. This is why results should be seen more as a collection of relevant views than one view that covers the whole geospatially related production processes.

At Statistics Finland the GSBPM is in use widely as a process framework for renewals of statistical production processes. However, it has not been used in describing geospatial related production processes. In the GEOSTAT 2 project Statistics Finland made preliminary testing from the geospatial point of view and tried to formulate the best work flow for testing in all project countries in order to achieve comparable results.

The Ten-Step instructions were assembled and they also describe the work flow that Statistics Finland followed (see Annex 3, Ten-Step instructions).

The first remark at Statistics Finland was that the model fits into geospatial data and processes if the abstraction level of the current model description is raised. The logical meaning of a particular phase or sub-process was widened and as a result geospatial processes and statistical processes could be fitted.

Statistics Sweden evaluated the GSBPM by reflecting it to their national statistical production process model (which is similar to the GSBPM but older). Possible gaps between the national model and the GSBPM and also gaps in both models were examined from the geospatial point of view.

As Statistics Finland, Statistics Sweden also found the current structure of the GSBPM sufficient to cover the geospatial production. However, according to Statistics Sweden’s point of view the GSBPM needs substantial development relating to semantics and terminology of geospatial elements provided in the descriptions.

At Statistics Austria the GSBPM is in use increasingly in developing statistical production processes and structuring activities. Statistics Austria’s first impressions have been that the GSBPM could cover the geospatial production but the text amendments of geospatial elements would be needed. Even this first examination produces some very significant development needs concerning core data in geospatial statistical production.

INSEE (Statistics France) values the GSBPM as a powerful tool to implement strategic objectives into statistical production. INSEE sees that maintaining and updating data and ensuring the consistency of geospatial data should be added to the GSBPM. The terminology should be opened and defined before further development of the model is possible. INSEE sees that geospatial related sub-processes should be added as separate sub-processes.

At Statistics Portugal the national process model is to be redefined and developed to cover the GSBPM. The geographical references have been part of the new approach and will be part of the future redefining of the model. So far, Statistics Portugal has identified the GSBPM sub-processes where geospatial elements are to be seen.

At Statistics Poland a comparison between the GSBPM and the national process model has also been carried out. The practical implementation of relevant processes relating to spatial data and mapping them with GSBPM model showed that the important areas are not included in the model, although they are important and necessary in the actual production process. By indicating these areas, the potential shortcomings and imperfections of the GSBPM model were diagnosed. The analysis revealed shortages of the GSBPM model, which essentially concerned statistical data spatialisation aspects from the stage of designing the data collection, geocoding, analysis and providing spatial characteristics of statistical products. Statistics Poland suggests that four new sub-processes should be added to the GSBPM.

8 Terminology

The terms and their definitions used to describe the setup and use of a point-based geocoding infrastructure in this report are mainly based on information in existing UN-GGIM reports, the Statistical-Spatial-Framework of the Australian Bureau of Statistics, INSPIRE, the European Statistical System and Wikipedia.

A draft compilation of statistical-geospatial concepts and terminology was prepared by Ekkehard Petri (Eurostat) in 2015 on behalf of the UN-GGIM Expert Group on the Integration of Statistical and Geospatial Information. The content of this proposal for a common statistical-geospatial terminology database is published on www.EFGS.info[40]. Please visit this website for further information regarding definitions and explanations of statistical-geospatial concepts not explicitly described in this report.

[40] http://www.efgs.info/information-base/introduction/terminology/

9 Conclusions

The need for a more systematic approach to location and geospatial information in the collection, processing and dissemination of statistics has been acknowledged by the EU and its Member States for almost two decades. An overall conclusion from the GEOSTAT 2 project is that location needs to be tightly and fully integrated into the statistical production process. A point-based foundation for statistics is the key condition that enables NSIs and the European Statistical System to increase relevance and timeliness of statistics but also to improve efficiency and flexibility of production.

The main business of moving location into the core of statistical production is to set up a consistent and flexible point-based geocoding infrastructure for statistics. In a basic sense, a point-based geocoding infrastructure for statistics can be understood as a production setting where a record holding X, Y (and Z) coordinates of a location, along with a unique identifier (Id), can be linked to a record with statistical or administrative data which belongs to this point. The characteristics of such an infrastructure encompass the following three generic principles:

- Use of high quality point-based location data, regularly updated with time stamps
- Geocoding of statistical unit, and related statistical information, at unit record level
- Use of standardised identifiers/geocodes to link unit record data with location data

In order to mainstream the use of geospatial information in the statistical production process and to design geospatial work flows that can serve multiple statistical domains, a broad understanding of the different roles that geospatial data may play in the statistical production is needed. The GEOSTAT 2 project advocates NSIs to use existing business models and frameworks, e.g. the GSBPM and the Global Statistical Geospatial Framework (SGF), as tools to start fleshing out how to set up, or improve, cross-product solutions for geospatial information suitable for their particular business conditions. The power of the GSBPM for structuring this process and support the communication between statisticians and geospatial experts has been proven successfully in a number of NSIs as part of this project.

Building on the process-based approach of the GSBPM, this report has provided operational advice on how to set up and maintain a point-based geocoding infrastructure. The key tasks are elaborated in detail in Chapter 6, but in essence the content can be expressed in the following recommendations:

- Make sure to obtain a good understanding of the expectations and requirements of the users, both externally and internally. Introduction of a point-based foundation for statistics will and should ideally have an impact on several domains of the statistical production chain (from survey design, collection methods, processing techniques, quality assessment, sampling methods to dissemination and confidentiality etc.). Hence, it is important to engage the organisation broadly to find out how and where a point-based infrastructure for statistics can improve business processes.
- Make a thorough assessment of location data sources that can be potentially used to set up a point-based geocoding infrastructure and decide on a reasonable level of quality needed. Fundamental and authoritative geospatial data from the National Spatial Data Infrastructure should be the first-hand option. Make sure that the provider or custodian has a consistent and systematic scheme of maintenance to keep the dataset updated and that data is

properly described in line with metadata standards or any other official data specifications. It can be expected that the INSPIRE directive along with the work pursued by UN GGIM: Europe have triggered, or will trigger, a gradual improvement of access to geospatial data in many Member States.

- Build formal working relationships with external producers of geospatial information (e.g. NMCAs) as to safeguard long-term provision of data. Cooperation between institutions should ideally rely on (formal) agreements or legislation, but the agreements themselves are no guarantee for a good, flexible and solid cooperation. Cooperation with producers or data custodians should entail establishment of feed-back routines for reporting and correction of errors found in geospatial information.
- Develop a strategy to manage the geospatial data streams in the best possible way. Such a strategy should address the human resources needed as well as the technical infrastructure allowing for efficient processing of geospatial data (software, data storage, computing capabilities etc.). The goal should be a flexible technical environment and to avoid duplication of data and smooth processes for maintenance.
- Develop routines on how to deal with the temporal aspects of data in order to obtain the best possible temporal cohesion between location data and statistical and administrative data to be geocoded.
- Develop routines for a uniform approach to geocoding, most notably describing workarounds to handle erroneous data (erroneous ids, missing information etc.). Such routines may include address validation tools, homogenisation of address information or interpolation of address location points.

This report describes the complete geocoding infrastructure and its tight integration into the production process as the goal that all NSIs should strive to achieve and maintain. The survey conducted by the project at half-term revealed that the situation in NSIs with regard to the above recommendations is diverse and ranges from a patchy, scattered implementation of some of the above elements to a fully consistent production set-up that already covers almost all aspects although improvements are always possible.

The creation of a geocoding infrastructure for statistics and its integration into the statistical production process does not absolutely require a big-bang approach and a complete redesign of enterprise architectures, production processes and legislation inside and outside NSIs. Small and stepwise improvements are possible and experts in NSIs will find many good examples along the production process in Annex 1 that contains a collection of national best practices for all phases of the GSBPM. Experts and managers in NSIs should look at these individual use cases as a source for inspiration for future improvement projects that suit best their national context. Integration of statistical and geospatial information is a cornerstone in the modernisation of official statistics. The GEOSTAT 2 project consortium wishes all colleagues a lot of success on this long but rewarding journey.

10 References

Australian Bureau of Statistics 2015. SSF Guidance Material – Geocoding Unit Record Data Using Address and Location. Second release. http://www.nss.gov.au/nss/home.NSF/pages/Statistical+Spatial+Framework+Guidance+Material/$File/Geocoding%20Unit%20Record%20Data.pdf

Benedetti, R, Piersimoni, F & P, Postiglione 2015. Sampling Spatial Units for Agricultural Surveys. Springer.

EFGS/GEOSTAT 1 2012. GEOSTAT 1A – Representing Census data in a European population grid. Final report. http://ec.europa.eu/eurostat/documents/4311134/4350174/ESSnet-project-GEOSTAT1A-final-report_0.pdf

EFGS/GEOSTAT 1 2013. Production procedures for a harmonised European Population Grid. Aggregation method.

EFGS/GEOSTAT 1 2014. GEOSTAT 1B Final report. https://ec.europa.eu/eurostat/cros/system/files/WP0-4_geostat1b-final-technical-report-v5.pdf

EFGS/GEOSTAT 2 2016. Spatializing statistics in the ESS – Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs. http://www.efgs.info/wp-content/uploads/geostat/2/GEOSTAT2-Spatialising-statistics-in-the-ESS-2015-002.pdf

ESS Vision 2020. http://ec.europa.eu/eurostat/documents/7330775/7339647/ESS+vision+2020+brochure/4baffcaa-9469-4372-b1ea-40784ca1db62

European Commission 2013. INSPIRE D2.8.III.2. Data Specification on Buildings – Guidelines.

European Commission 2009. INSPIRE D2.8.I.6. Data Specification on Cadastral Parcels – Guidelines.

Eurostat 2002. TANDEM GIS I – A (feasibility) study towards a common geographical base for statistics across the European Union. http://ec.europa.eu/eurostat/documents/3888793/5816633/KS-AN-02-001-EN.PDF/fffe0de6-009f-4abc-b5c7-d553c388583c

Eurostat 2015. Report from the task force on the integration of statistics and geospatial information. https://circabc.europa.eu/d/a/workspace/SpacesStore/fd349927-3c7d-435a-8b80-4ace48daf646/D_GIS_105%20GISCO-TF-Report-V_3.doc

Haussman, M, Maack, U & K, Trutzel 2015. Merging Statistics and Geospatial Information. The Urban Contribution to the European Spatial Data Infrastructure. A project of the KOSIS Association Urban Audit in cooperation with the KOSIS Association DUVA and KORIS 2014-2015. Excerpts of the Final Report on the Results of the Working Package 1. Georeferencing administrative Register.

UNECE 2013. Generic Statistical Information Model (GSIM): Specification (Version 1.1, December 2013) http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification

UNECE 2013. Generic Statistical Business Process Model GSBPM (Version 5.0, December 2013) http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

UNECE 2015. Common Statistical Production Architecture (Version 1.5, December 2015) http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.5

UNECE 2015. GAMSO Generic Activity Model for Statistical Organisations (Version 1.0: 1 March 2015) http://www1.unece.org/stat/platform/display/GAMSO/GAMSO+v1.0

UNECE 2016. In-depth review of developing geospatial information services based on official statistics. Note by the UK Office for National Statistics. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/bur/2016/February/02-In-depth_review_on_developing_geospatial_information_final.pdf

UNECE 2016. Statistical and Geospatial Information – an Australian perspective on challenges and opportunities. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2016/mtg/CES_21_Eng_G1602515.pdf

UNECE 2016. United Nations initiative on Global Geospatial Information Management (UN-GGIM) – All about connections. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2016/mtg/ECE_CES_2016_20-1602011E.pdf

United Nations Expert Group on the Integration of Statistical and Geospatial Information 2016. Background Document on Proposal for a Global Statistical Geospatial Framework (Advanced Draft as of 28/07/2016). http://ggim.un.org/docs/meetings/GGIM6/Background-Paper-Proposal-for-a-global-statistical-geospatial-framework.pdf

United Nations Sustainable Development Goals (SDGs). http://unstats.un.org/sdgs/

UN-GGIM Europe 2016. Core Data Scope. Working Group A – the First Deliverable of Task 1.a. http://un-ggim-europe.org/sites/default/files/UN-GGIM-Europe%20WGA%20Core_Data_Scope-v1.2.pdf