GeoStat

GSGF Europe: Guidance and Analytical Tools for Implementation of GSGF Europe Reference Architecture

GEOSTAT 4

eurostat

Summary

As a continuation of the GSGF reference architecture, supporting material for its actual implementation has been designed. It aims to guide stakeholders in evaluating their own context in comparison with the GSGF Europe specification, identifying components that are already available, and those to be built. The supporting material aims at highlighting specific and non-shareable technical implementations of statistical processes, in order to replace them with shared and generic solutions, in accordance with international standards. Two kinds of supporting materials are proposed: schematic representations of the way data are transformed within geospatial statistical processes, with different trajectories illustrating different business and technical contexts; a first trial of analytical grids for a selection of sub-processes to help assessing the maturity level, from an architectural point of view, at each stage of the statistical geospatial process. Stakeholders are invited to use these tools as a help to build transversal knowledge of the processes currently at work, and to establish an achievable target, allowing progress to be made in accordance with the principles of the GSGF.

Table of contents

1. Introduction

Implementing a reference architecture depends upon a variety of institutional, methodological and technical factors, that may lead to different pathways for each country and organisation that are willing to endorse such a general framework. It is thus necessary to help actors in their discovery of the reference architecture, then in the way of adapting it to their national context, depending on their business environments and technologies. A template solution aligned on the GSGF Europe has been defined with a complete panorama of the different dimensions to be covered in a reference architecture for the production of geospatial statistics. These dimensions are

- a mapping of actors
- a description of the processes and activities based on the GeoGSBPM, an adaptation of the GSBPM to the statistical geospatial production processes;
- a conceptual model linking geographic data and statistical data;
- an inventory of services specific to the production of geospatial statistics.

The reference architecture embodies best practices, typically suggesting the optimal method based on commonalities and shared resources, as "*Enterprise architecture helps to remove silos, improves collaboration across an organisation and ensures that the technology is aligned to the business needs*." (see "GEOSTAT 4: GSGF Europe reference architecture" document). As a continuation, supporting material for the actual implementation, called an Implementation Companion of the GEOSTAT Reference Architecture has been designed. It aims as an analytical tool to guide stakeholders in evaluating their own context in comparison with the GSGF Europe specification, identifying components that are already available, and those to be built, both on a Business level and Service Level.

The Implementation Companion gives practical means to cross the different dimensions in a «contextualized collage ». These materials can be used by stakeholders to help them in choosing priority investments (on institutional, organisational or technical topics) and consistent solutions in respect with actual practices. In that sense, the Implementation Companion of the Reference Architecture can be seen as part of a maturity model, to be used for evaluating and measuring organizational progress and success in adoption and implementation of the GSGF Europe. It is made up of two elements: an analysis grid in the form of a spreadsheet articulated by process phases, and schematic representations built around the production cycle of a statistical geospatial content. This document aims to introduce this Implementation Companion and present its underlying features.

2. CSPA and GeoGSBPM as related frameworks

The architecture developed for the production processes of geospatial statistics is based on the Common statistical production architecture (CSPA), and its different layers. The CSPA is at the very heart of architectural effort to standardize and industrialize statistical processes, as it defines « a set of agreed common principles and definitions designed to promote greater interoperability within and between different stakeholders that make up the (statistical) industry" (see "GEOSTAT 4: GSGF and Other Frameworks" document"). The CPSA promotes a service-oriented architecture based on standardization at each level: the business level, the information level, and the technical level.

The modernisation blueprint in the CSPA framework basically includes three different highlights, that are embedded in the Implementation Companion and its analytical grid.

1. a common description of statistical pipelines, facilitated by common grammar for describing their phases.
2. demise of specific and non-shareable technical implementations of statistical processes, and coordinated development of generic services
3. re-building of statistical pipelines as a plug-and-play assembly of shared or reused services, based on interoperability.

The first highlight requires the use of a common grammar which delivers a universal breakdown of statistical processes in main phases and sub-processes. The GSBPM is nowadays widely used throughout the statistical community so as to speak a unique language in describing the main steps of a process. For that reason, the Implementation Companion is also based on such a grammar, and the analytical grid includes one tab for a selection of key GSBPM sub-processes. Some phases are left aside, because they do not include geospatial specificities or have no link with the GSGF Europe. Going beyond the general GSBPM framework, the GeoGSBPM takes into account the specific aspects of the production of geospatial statistics. For each selected sub-process, geospatial-related activities resulting from the GeoGSBPM (also used as is in GSGF reference architecture) are recalled and summarized in the analytical grid. Some may be added if relevant.

The second highlight promotes the establishment of a set of generic services, which purpose is to address, for each of them, dedicated part of the statistical geospatial pipelines. The standardized breakdown of the statistical process into phases and sub-phases facilitates the design of the required building blocks. Once assembled and deployed for a given purpose, building blocks bring needed services to life. Compliance with the CSPA reinforces the ability to build and share internationally, -or reuse existing- building blocks. But this objective can only be achieved in organizations with a sufficiently advanced context, which is the rationale behind the analytical grid as a tool to assess the services that are already implemented, and those which could be developed.

The third highlight invites to design pipelines as a plug-and-play assembly of shared or reused services. Regarding the statistical geospatial pipelines and the GSGF Europe, this goal implies a supplementary requirement, as in fact two different kinds of processes – and actors - are at stake: "In the GSGF, the basis is that statistical processes are very dependent on the geospatial processes, which are usually carried out by NGIAs and aim to produce geospatial reference data and tools used by a wide community of data producers" (see "GEOSTAT 4 : GSGF Europe" document). This specificity raises two issues: the cooperation of actors to facilitate and equip data exchanges and interoperability. The description of actual and target processes within the Implementation Companion thus invites to tag geospatial and statistical "facets" for each component, with a special focus on three elements: reference geospatial data, shared repositories and shared services.

The requirement for interoperability is not limited to national processes, it also concerns international exchanges. Hence, there is a need to promote standards and technical solutions at a supra-national level to harmonize geospatial data from one country to another and in line with European regulations in this area, which is usually considered within the analytical grid as the most advanced level of compliance with GSGF principles.

3. The data journey through geo-enabled statistical pipelines

In order to guide them in the implementation of the GSGF Europe, stakeholders are first invited to describe the current flows of information which is progressively enriched on spatial and statistical dimensions. This "datacentric" description of actual processes is to be covered using the common grammar provided by the GeoGSBPM and, whenever possible, using the GSIM and the conceptual model included in the GSGF Reference Architecture.

The work carried out on the process phases and sub-phases aims to define more precisely the services used in actual processes, and those which could be built in the future to be in line with the GSGF principles and recommendations. Most of these services are identified in the service map drawn up as part of the architectural work. A second objective is to identify, over the course of the "data journey", the capitalization that is carried out on the processed information, by feeding data repositories -and vice versa, by feeding on these data repositories. Build and use of data repositories is a key aspect so as to ensure consistency between several processes that have to be based on common references (for instance, for geocoding), or to avoid re-carrying out treatments which results can be reused. A third objective is to understand, for each service or for each data repository, whether the latter is primarily under the responsibility of the geospatial community - and therefore constitutes part of the National Spatial Data Infrastructure - or falls under the responsibility of the statistical community - and should be included, in a similar way, within a National Statistical Data Infrastructure. A final objective is to identify the sub-phases of statistical processes which are particularly concerned by interactions with geospatial processes - in other words, to highlight exchanges to be structured between the two types of infrastructures, a key point to agree on governance principles for these commons.

Describing these “data journeys” aims to provide an inventory of the current situation, then, guidelines to define a realistic trajectory for implementing GSGF compliant processes, taking into account legacy and context. The implementation companion includes an example of Data journey description, in the form of a schematic representation with different stages, services and repositories that are used. In this example, several paths are illustrated, in order to take into account for different trajectories, the implementation of which will depend on prerequisites linked to the national context. Finally, reference domains are suggested for services, data repositories and final delivers, in association with a colour (green or purple) which refers to the geospatial or statistical community as "product owner"

Actually, three main factors determine the context to be taken into account:

- a factor relating to the actors: the institutional context in which the processes take place, the level of cooperation between the actors of the geospatial data production processes on the one hand and the producers of statistical data on the other hand, and the existence or not of shared governance.
- a technical factor relating to IT architectures: the level of tool standardization achieved in comparison with the "optimal" level of standardization for the production and transformation of data. This standardization is essential for the building of shared services for several processes, and brings technical challenges to implement working interfaces that make a bridge between different information systems, for example between a National Statistical Institute and a National Mapping Agency.
- a semantic factor which corresponds to a specific dimension of interoperability: the consistency in the definition of concepts and entities. Semantic consistency is critical to make the link between the stages of production and those of dissemination, which is even more complex regarding statistical geospatial data, based on separate but very intertwined processes.

The data journey example description included within the Implementation Companion is inspired by the Common Statistical Data Architecture (CSDA). CSDA shows organisations how to structure their processes and systems for efficient and effective management in a “datacentric” perspective, from the external sources through the internal storage and processing up to the dissemination of the statistical end-products (see “GEOSTAT 4: GSGF and Other Frameworks” document).

Two levels of data journey description are provided as a first canvas to be adapted in local context: either a macro view covering a GSBPM phase as a whole, or a micro view looking deeper into GSBPM sub-processes. Only a set of phases and sub-processes are covered, as some of them are not appropriate to describe the actual data journey1 – till they can be partly discussed in another part of the Implementation Companion, through the analytical grid.

Canvas is available in (Document_1_Implementation_Companion_Schema_Canvas)

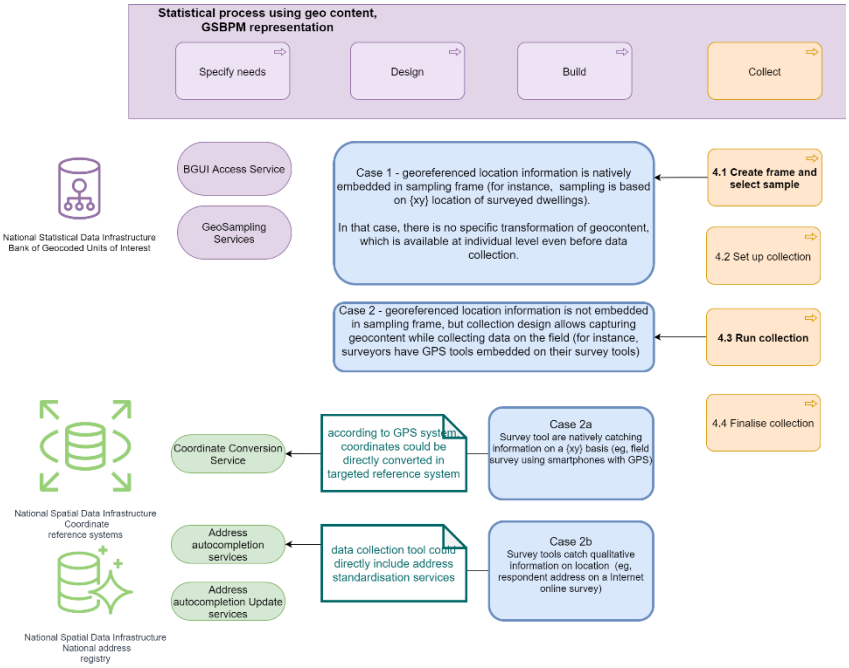

Figure 1: Alternative data journeys described at a macro level, for "Collect" phase

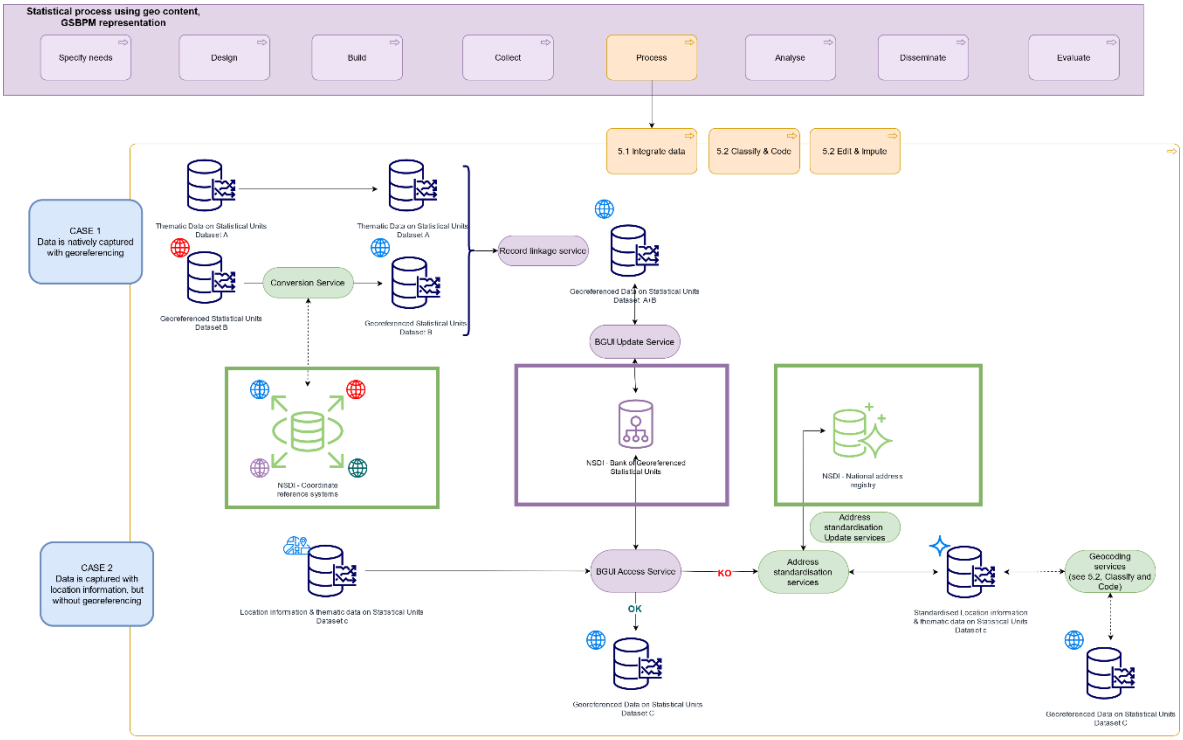

Figure 2: Alternative data journeys described at a micro level, for a selection of "Process" sub-processes

4. Going towards self-assessment of maturity level

In addition to data journeys' diagrams, a first trial of analytical grids for a selection of statistical process sub-phases are proposed within the Implementation Companion. The analytical grids aim to help assessing the maturity level, from an architectural point of view, at each stage of the statistical geospatial process. They show how the context can be more or less favourable depending on which part of statistical geospatial processes we are looking at - as it might be difficult to assess the process at once, as a whole. For example, a national context may be particularly advanced in terms of common governance between statistical and geospatial communities, with a shared definition of semantics, but at the same time it can lack IT systems to integrate these elements as common services. In another case, tools can be common on certain production stages (for example, geocoding), but not dissemination stages (for example, distribution via harmonized catalogues of datasets), or the other way.

To characterise this context and its cross-cutting dimension, it is proposed that the maturity level be evaluated on requirements specific to selected sub-phases. Each sub-process may be more relevant to the institutional side, the technical side of IT system, the semantic side on concepts, or all of them, for both the geospatial production and statistical production. Indicators included in analytical grids are expressed as examples and a very first step to design a complete maturity analysis – which might be the topic of future work forward from Geostat4 ESSNet. These indicators relate to the existence (or not) of statistical and geospatial services, and their level of compliance with the GSGF principles, based on the applicable standards (for example, OGC standards) and on the components developed in the reference architecture (list of actors, list of services, etc.)

The analysis grid proposed in the Companion provides a simple and generic framework for verifying, for each sub-phase selected, the current maturity level and the attainable objective. It has several elements:

- the general description of the sub-phase, according to the GeoGSBPM, listing the activities that have a geospatial component
- the different types of services that can be mobilized, in order to spot those that exist in the national context
- the principles of the GSGF which are linked to the sub-phase
- if necessary, quality criteria can also be added.

Each selected sub-phase is presented on a dedicated tab inside the summary spreadsheet which gathers information. Analytical grids are available in (Document_2_Analytical_Grid_Spreadsheet)

Within the analytical grid, two dimensions deserve to be taken into account.

One is the result of institutional cooperation, through existence of shared tools, used jointly by production and dissemination processes of geospatial data on the one hand, and statistical geospatial data processes on the other. Data repositories storing reusable and reference information, and their related access and update services, are good examples of these items strongly dependent on governance issues.

The other is specific to the statistical production architecture and its Information System components: the compliance level with CSPA specifications. The more the architectural principles are respected, the higher the level of adoption of international standards, the more the GSGF reference architecture implementation will be facilitated.

Sub-phase	Description	Source	Comment
5.1 Integrate Data	This sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined. The input data can be from a mixture of external or internal sources, and a variety of the collection instruments, including extracts of administrative and other non-statistical data sources. Administrative data or other non-statistical sources of data can substitute for all or some of the variables directly collected from survey. This sub-process also includes harmonising or creating new figures that agree between sources of data. The result is a set of linked data. Data integration can include: • Combining data from multiple sources, as part of the creation of integrated statistics such as national accounts; • Combining geospatial data and statistical data or other non-statistical data; • Data pooling, with the aim of increasing the effective number of observations of some phenomena; • Matching or record linkage routines, with the aim of linking micro or macro data from different sources; • Data fusion - integration followed by reduction or replacement; • Prioritising, when two or more sources contain data for the same variable, with potentially different values.	GSBPM 5.1 (79)	Data integration may take place at any point in this phase, before or after any of the other sub-processes. There may also be several instances of data integration in any statistical business process. **Following integration, depending on data protection requirements, data may be de-identified, that is stripped of identifiers such as name and address, to help to protect confidentiality.**
	Activities		
	Combinie geospatial data and statistical data or other non-statistical data	Unece (70 & 71)	When combining geospatial data with other data, the geospatial units used in the datasets might be different (e.g. grid in population data and administrative boundaries in agricultural data), matching strategy should be consistently applied and any non-matching should be documented with quality measures as developed in Phase 2 Design.
	Use of location as a matching key variable	Unece (72)	Geospatial information (e.g. address, coordinate, geographical names) can also play an important role in bringing together information from various domains by enabling integration of datasets from different sources using the location information as a matching key variable (e.g. integrating administrative data with survey data using address or postal code in both datasets). To ensure the quality of the integration, standardising the geospatial information in the different datasets is critical. This standardisation would normally take place before the integration of datasets and can be done through, for example, matching location information in the datasets with centralised standard system (e.g. address matching, geocoding) which can also allow the dataset to use various additional geographical information within the address registry or geocode database. In the absence of such system, organisations may rely on other ways of reference location (e.g. GPS coordinate), alternative source (e.g. address registry from utility provider) or higher-level geography (e.g. regional boundary).
Maturity level			
Level 1	Integrated Datasets cannot be merged at individual levels due to discrepancies in geospatial coding or units between datasets (for instance, each source uses a specific geo specification, without having a mapping between different specifications). Matching can only be done at aggregated levels		
Level 2a	Integrated Datasets use their own specific geo specification, but a « conversion » service can be used to refer to a common geospatial reference and linked datasets at individual levels (eg : coordinate transformation between different geodetic coordinate reference systems).		
Level 2b	Integrated Datasets use a common geospatial system. Data Integration process uses standardisation services so as to normalize geo information and maximize matching results (eg : address standardisation service based on a unique national address registry)		
Level 3	The integrated data are confronted with a unified source of knowledge on the geocoding of statistical units, making it possible to associate a geocoding with it and, if necessary, to update the unified source of knowledge.		
Related Services			
Conversion service			
Level 1	None		
Level 2a	Coordinate transformation		
Level 2b	(i) National address registry (NAR) (ii) NAR extraction/consultation service (iii) NAR standardisation service		
Level 3	(i) Bank of Geocoded Units Of Interests (BGUI) (ii) BGUI extraction/consultation service (iii) BGUI update service		
Merging service			
Level 1	None		
Level 2	Ad-hoc merging services		Proprietary services built up for specific purpose
Level 3	Open Stardard Table join services		Shared service for joining geospatial data and statistical data, e.g. TJS
Related GSGF Principles			
Principle 2	Geocoded record unit data in a data management environment	Requirement 2.2	Store location only once
Principle 1	Use of fundamental geospatial infrastructure and geocoding of statistical information	Requirement 1.1	Use data from NSDI

Figure 3. Analysis Grid illustrated on "Integrate Data" GSBPM Sub-process, with different maturity levels and two services (conversion service and merging service).