Data Integration Methods

Analysis of future trends in geospatial data capture, creation, maintenance and management and recommendations for amplified use of good practices

UN-GGIM: Europe | Working Group on Data Integration

Version 1.0

2021-11-22

UN-GGIM: EUROPE | UNITED NATIONS COMMITTEE OF EXPERTS ON GLOBAL GEOSPATIAL INFORMATION MANAGEMENT

CONTENT

FIGURES

0 Executive summary

Data integration methods exist today; they are applied and well documented. Moreover, the technology is rapidly evolving. The objective of this document is: to summarize the situation on data integration in Europe mainly between geospatial and statistical data, to illuminate the data integration barriers from a technical point of view, and to provide a global and cross-domain perspective with recommendations for amplified use of good practices.

This will be an incentive for discovering, encouraging and initiating the use of data integration methods and will facilitate the decision-making process of the countries, for seeking and governing new approaches in their data capture, creation, maintenance and management. Therefore, the document addresses senior managers and senior consultants mainly affiliated to Geospatial Agencies (National Mapping and Cadastral Agencies, NMCAs) and National Statistical Offices (NSO). However, as a basic understanding of the technical foundation, it is a prerequisite for considering the implementation of the data integration processes, the technical foundations of data integration methods and technical considerations are introduced in this document, too.

In order to analyse the different data integration methods and to provide further insights, the 'UN-GGIM: Europe-Working Group on Data Integration' observed national examples in the European region. The aim was to identify the most significant methods for further recommendation. A limited set of methods – highlighting the most relevant aspects and impact on data integration – is extracted and explained further below. It is worth noting that some examples of methods refer explicitly to the integration of geospatial data and statistics, while others provide a broader perspective for cross-domain data integration methods coming from Libraries or Digital Humanities. Anyway, these methods addressing only the integration of geospatial data and statistics and their ensuing recommendations might be used as a blueprint for other domains.

The document underlines that the 'Linked Data' method should be seen as the key enabler for data integration; it is a key data integration method and therefore a good starting point to further elaborate the perspectives within the geospatial and the statistical community, collaborating within UN-GGIM and in the European regional committee UN-GGIM: Europe.

Data integration shall also be understood here as an essential part of a geospatial knowledge infrastructure aiming at common data spaces. The European political program and the European Green deal – which establishes a green transformation in the light of the UN Agenda 2030 for Sustainable Development Goals (SDG) – have an impact on the further development of data integration methods; besides the existence of interoperable standardised spatial data infrastructure like INSPIRE, this shows the limitations affecting the effective use of it and, geospatial data infrastructures should evolve to a broader context. Spatial is not special anymore; the spatial data infrastructures and the geospatial data have to adapt to an integration into the European Data Spaces.

The concluding part of the document raises recommendations for an amplified use of good practices enabling a successful data integration process in a technical perspective. A call for action in Europe should address the recommendations put forward in this document.

These recommendations can be differentiated into those addressing the **(1) linking of geospatial and statistical data (Linked Data)** and those acting on the **(2) knowledge infrastructure and on the data**

infrastructure, which basically provide the suitable conditions for facilitating the data integration processes. All recommendations are listed and explained with more details in chapter ‘Recommendations’, but can be summarized as follows;

(1) The recommendations for successfully linking geospatial data and statistical data (Linked Data) are:

- to define and implement valid Persistent Identifiers (PID) across domains;
- to agree upon common definitions for fundamental geographies and linked data and for ontologies (definitions of terms and relations);
- to make geospatial/statistical data interoperable, simple, and of good quality in order to enable trustful and successful data integration processes throughout European data spaces;
- to develop and to implement sustainable and automated data integration processes;
- to develop and to implement standardized and open APIs for a smart geospatial data provision.

(2) The recommendations for building the knowledge infrastructure and the geospatial data infrastructure are:

- to invest resources and capacity building into a cross-domain Geospatial Knowledge Infrastructure (GKI) aiming at common data spaces;
- to modernize the current National Spatial Data Infrastructures (NSDI) towards GKIs.

1 Introduction

In a changing world with new technologies, diverse data sources and user demands moving towards a real-time delivery of knowledge, the focus of this document prepared by the 'UN-GGIM Europe-Working Group on Data Integration' is on highlighting different technical methods for a more efficient data management and data integration pertaining to geospatial and statistical data.

Therefore, several issues have to be considered. Data integration follows the intention to combine – or to link – information from different sources and domains in order to receive comprehensive insights; differing data structures, organisational issues and legal constraints are well known barriers to prevent data integration; even if all those issues are solved, which means, amongst others, that a common data structure exists, organisations provide harmonised data and the licenses are creative commons, data integration will not work out of the box or in an automated way yet.

One main reason is that the information derived from data is context dependent, which is known as semantic dimension; 'wood', for example, could be an area provided as geometric polygon that represents an area for touristic leisure for one community, whereas it is avalanche protection or an economic forestry for others. This semantic dimension can be unambiguously formalized with common vocabularies, relations between these vocabularies and ontologies that describe the schema of relations – how features are connected.

Data integration is also an integral part of the (geospatial) knowledge infrastructure; with the help of persistent repositories – sustainable available registries and information sources – and Persistent Unique Identifiers (PID), a common, cross-domain and usable geospatial knowledge infrastructure can be established.

The technology of data integration methods and of data sources is also rapidly evolving like sensor observation from satellites, airplanes, cars, mobile phones. This can be used to generate new insights and to have own storage and usage; examples are: Data Lake, Big Data, Analysis Ready Data, Data Cubes, and so forth. However, the working group made a pragmatic selection based on the methods applied within the geospatial and the statistical community and left further topics out purely due to space constrains. Anyway, these methods addressing only the integration of geospatial data and statistics, and their ensuing recommendations may be used as a blueprint for other domains.

The document is structured in two main parts; the first part addresses the rationale, the political and technical context and the second part addresses the technical foundations of data integration methods based on national examples and the ensuing recommendations.

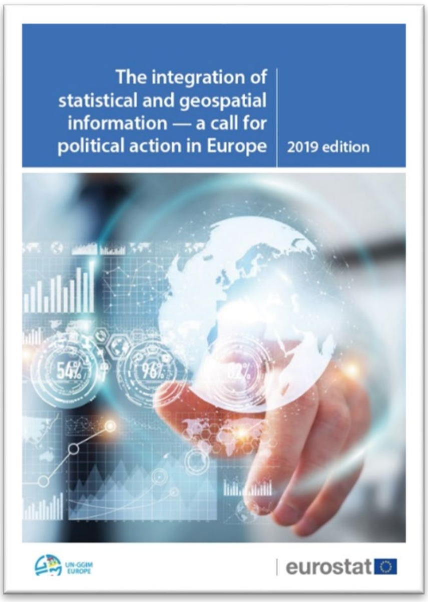
In the first part of the document, the chapter 'Rationale' briefly describes the previous work done in the working group analysing and describing high-level recommendations for the integration of geospatial data and statistics; the need for a more practical implementation of data integration is highlighted. The chapter 'Political and Technical Context' outlines the requirements for further development of data integration methods, and their impacts in the light of the European political program, the European Green Deal and the UN Agenda 2030 for Sustainable Development Goals (SDG), as well as in the light of current spatial data infrastructures like INSPIRE.

In the second part of the document, the chapter 'National Examples' briefly describes the methodology for assessing national examples of data integration methods collected among the Working Group members. The chapter 'Data Integration Methods' describes the most frequently used methods for integrating geospatial and statistical data in the examples, and highlights the most relevant aspects on

data integration. The chapter ‘Data Integration Processes’ highlights several interacting key processes that shall be considered to run sustainable data integration processes.
The document concludes with chapters on ‘Recommendations’ and ‘Future Perspectives’. Additionally, glossary and annex are provided.

2 Rationale

Taking a step back, the regional committee UN-GGIM Europe has identified and communicated a list of issues that need to be addressed nationally and internationally in order to enable statistical and geospatial national agencies to deliver the most suitable data, to meet new user demands, and to give policy makers the power of evidence-based decision making. The results have been presented as recommendations for coordinated action in all European countries through a ‘Policy Outreach Paper’ published with the title ‘The integration of statistical and geospatial information —a call for political action in Europe’[1] in 2019 as part of the work plan 2017-2019 for the UN-GGIM Europe Working Group on Data Integration (see Figure 1).



The integration of statistical and geospatial information — a call for political action in Europe

2019 edition

UN-GGIM EUROPE

eurostat

FIGURE 1: THE ‘POLICY OUTREACH PAPER’

These recommendations were reviewed by the Member States giving comments and feedback. From the review[2] it can be inferred that the recommendations mentioned in the ‘Policy Outreach Paper’ are high-level and that practical implementations of data integration need to be further elaborated. Timely

[1] https://un-ggim-europe.org/wp-content/uploads/2019/09/KS0319423ENC_new.pdf

[2] https://un-ggim-europe.org/wp-content/uploads/2020/07/2020-06-30_UNGGIM-Europe_ReportEvaluationPolicyPaper_v1.0.pdf

and reliable information from a multiplicity of institutions are required to address current—e.g. COVID19—and future challenges—e.g. Climate Change, Sustainable Development Goals, Census 2021.

A lot of work activities have already been done by the United Nations Committee of Expert on the Global Geospatial Information Management (CoE UN-GGIM), the United Nations Expert Group on the integration of Statistical and Geospatial Information (UN EG-ISGI)[3] and the GEOSTAT[4] projects funded by Eurostat. These work activities include, amongst others, the Global Fundamental Geospatial Data Themes [5] considered as fundamental for strengthening a country's geospatial information infrastructure and the European implementation guide (GSGF Europe)[6] intended as an enhancement to the global guidance, addressing the regional specifics of Europe. All these results are important and referenced accordingly.

[3] http://ggim.un.org/UNGGIM-expert-group/

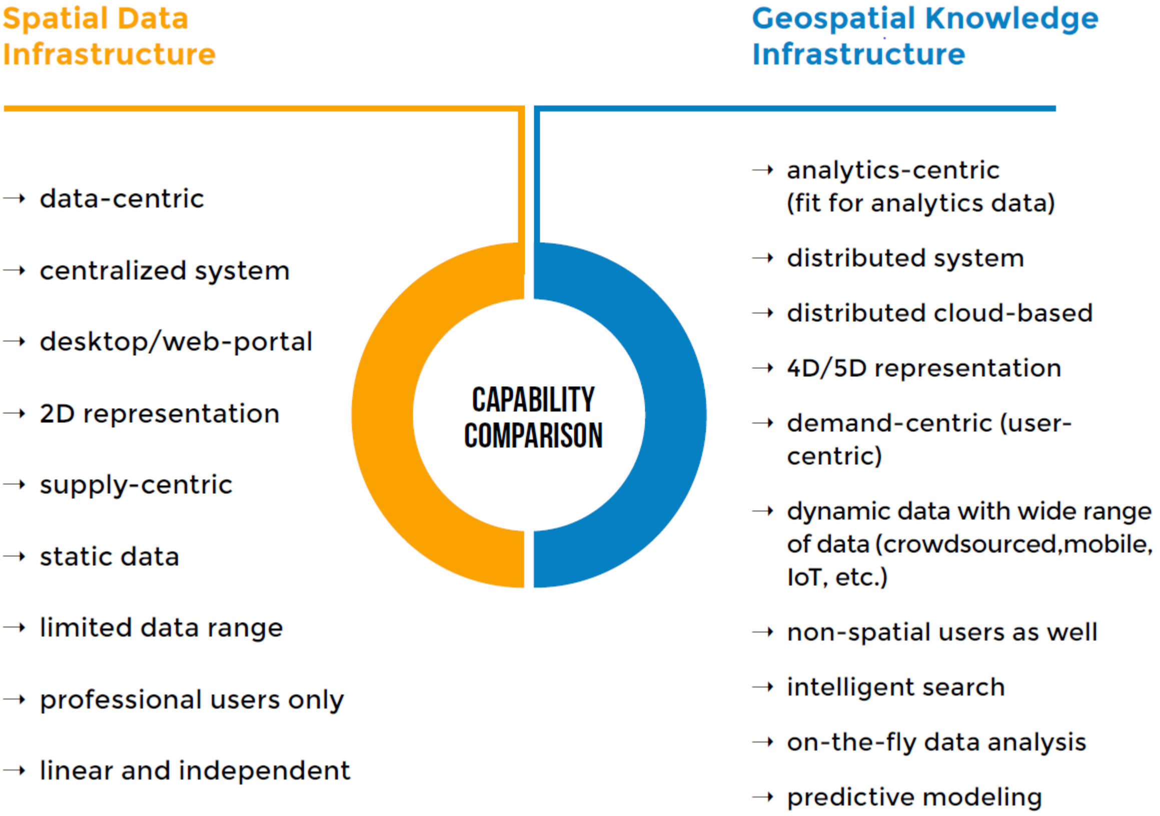
[4] https://www.efgs.info/geostat/

[5] UN-GGIM Global Fundamental Geospatial Data Themes: This publication was presented at the seventh session of the Committee of Experts UN-GGIM which adopted the proposed minimum list of global fundamental geospatial data themes. At its eighth session, the detailed theme descriptions were presented. An interactive presentation of the themes, developed by the UN-GGIM Secretariat, acts as a companion piece to this publication: https://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf

[6] European implementation guide (GSGF Europe): The implementation guide was proposed by the GEOSTAT 3 project. It aims to be more specific on the "how" to provide regional guidance on what elements should be available in countries to represent a meaningful GSGF: https://www.efgs.info/wp-content/uploads/geostat/3/GEOSTAT3_GSGF_EuropeanImplementationGuide_v1.0.pdf

3 Political and Technical Context

The Geospatial Knowledge Infrastructure discussion document[7] mentions the impact of the so-called 4th Industrial Revolution (4IR); "4IR is changing the world and the geospatial sector as well as the statistical sector must change with it. New technologies, varied data sources and user demand will require a next-generation geospatial infrastructure that embraces automation, dynamicity and real-time delivery of knowledge".



Spatial Data Infrastructure
Geospatial Knowledge Infrastructure
CAPABILITY COMPARISON
→ data-centric
→ centralized system
→ desktop/web-portal
→ 2D representation
→ supply-centric
→ static data
→ limited data range
→ professional users only
→ linear and independent
→ analytics-centric (fit for analytics data)
→ distributed system
→ distributed cloud-based
→ 4D/5D representation
→ demand-centric (user-centric)
→ dynamic data with wide range of data (crowdsourced,mobile, IoT, etc.)
→ non-spatial users as well
→ intelligent search
→ on-the-fly data analysis
→ predictive modeling

FIGURE 2: SPATIAL DATA INFRASTRUCTURE VS. GEOSPATIAL KNOWLEDGE INFRASTRUCTURE[8]

Figure 2 reflects the development from data to information to knowledge and finally to wisdom. This development emphasizes the changes from a spatial data infrastructure – here, the collection of data – to a spatial knowledge infrastructure – considered as the established meaning within the information, which is derived from data.

In Europe, geospatial datasets can be accessed through national, regional and local spatial data infrastructures (SDIs) for almost all administrative levels in many countries. The INSPIRE[9] initiative has

[7] https://geospatialmedia.net/pdf/GKI-Discussion-Document-Ver1.0.pdf

[8] Advancing Role of Geospatial Knowledge Infrastructure in World Economy, Society and Environment, page 20, https://geospatialmedia.net/pdf/GKI-Discussion-Document-Ver1.0.pdf

[9] Infrastructure for Spatial Information in Europe (INSPIRE) https://inspire.ec.europa.eu/

the ambition to standardise semantic and technical interoperability of geospatial datasets. Today, international agreed standards exist for an interoperable provision and processing of geospatial data. Legal frameworks are in place to integrate and use open geospatial data. However, many users are not aware of authoritative geospatial data provided by national governmental agencies or are prevented from using authoritative data due to their complex procedures for usage. The easy use of Volunteered Geographic Information (VGI) – like Open Street Map – or geospatial data centrally provided by the major web companies is common.

The reasons are manifold, amongst others:

- the procedures for using different data sources are still too complex;
- the data models for geospatial data provided through SDIs are considered too hierarchical or are focused on specific usage;
- the data exchange formats are specialised;
- the datasets mostly cannot be searched via commonly used search engines (like Google);
- the datasets are designed to support stationary applications like GIS software or tools only;
- the technology used to implement the SDI is only understandable for specialised developers within the geospatial community.

In a nutshell, SDIs are functioning well for highly skilled users within the geo community and do provide geospatial data, but most common users, like citizens or even officials-in-charge, prefer alternative geospatial data access provided via simple and easy-to-handle access points on the web[10].

The 2017 document 'Spatial Data on the Web Best Practices' describes in chapter 11 *"Why are traditional Spatial Data Infrastructures not enough?"*[11] the obstacles and possible solutions for the provision and publication of geospatial data on the web. The main recommendations include the use of persistent global Unique Resource Identifiers (URI) for 'Spatial Things' and describe the prerequisites for making geospatial data searchable via common search engines and for linking data to create the web of data. The Application Programming Interfaces (APIs)[12] shall be used as simple and easy-to-handle interfaces to access the web. However, recent and current developments by Open Geospatial Consortium (OGC) are renewing widely and profoundly geospatial services with a series of OGC API standards, partly already implemented in the most advanced NMCAs, changing the scene rapidly in the near future with a focus to bring geospatial capabilities fully operational on the web.

From simple visualisations to sophisticated interactive tools, there is a growing reliance on geospatial data by applications accessing web resources, so the need for sustainable and modern data infrastructures is obvious. In addition, the data volume offered by the private sector: sensor data in real-time, VGI, satellite and remote sensing data for free – e.g. via Copernicus –, is continuously increasing on the web.

[10] Taken from a presentation by Markus Seifert, Germany, https://www.frankfurt-university.de/fileadmin/standard/Studium/Studiengaenge/Fb_1/Bachelor-Studiengaenge/GeKo_BA/Dokumente/Geod_Kolloquium/Fb1_GeKo_DVW_Geod_Kolloquium_2020-11-12_Next-Gen_GDI_Seifert.pdf - in German

[11] Spatial Data on the Web - Best Practices, W3C Working Group Note, 28 September 2017, https://www.w3.org/TR/sdw-bp/

[12] https://ogcapi.ogc.org/

The European Data Strategy[13] – shaping Europe's digital future – is one example to create a single market for data – not only geospatial – to make the European Union globally more competitive and to enable innovative processes, products and services. In addition, the European Green Deal, which "[…] aims to make Europe the first climate-neutral continent by 2050"[14], requires strong support by geospatial data.

Accessible and interoperable data are at the heart of data-driven innovation. These data, combined with a digital infrastructure and artificial intelligence solutions, facilitate evidence-based decisions and expand the capacity to understand and tackle environmental challenges. The European Commission (EC) will support efforts to unlock the full benefits of the digital transformation to support the ecological transition. To do this, the EC will bring together European scientific and industrial excellence to develop a very high precision digital model of the earth. Interoperability (semantical) between so-called 'data spaces' is key!

The EU strives to become an attractive, secure and dynamic data economy, amongst others, by "[…] investing in the next generation standards, tools and infrastructures to store and process data" and by "[…] pooling European data in key sectors, with EU-wide common and interoperable data spaces."[15]

In order to achieve these technically related objectives, a modern and flexible technical framework has to be developed. Both the availability of competent API developers and well-educated users are crucial, besides the high-performance requirements for the web applications consuming the APIs.

Therefore, the SDI designers and architects have to realise that spatial is not special anymore and that SDIs and the geospatial data have to be adapted for the integration into the European Data Spaces (see Figure 3).

[13] A European Strategy for Data, https://ec.europa.eu/digital-single-market/en/european-strategy-data

[14] The European Green Deal, https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

[15] https://ec.europa.eu/eip/ageing/news/european-data-strategy_en.html

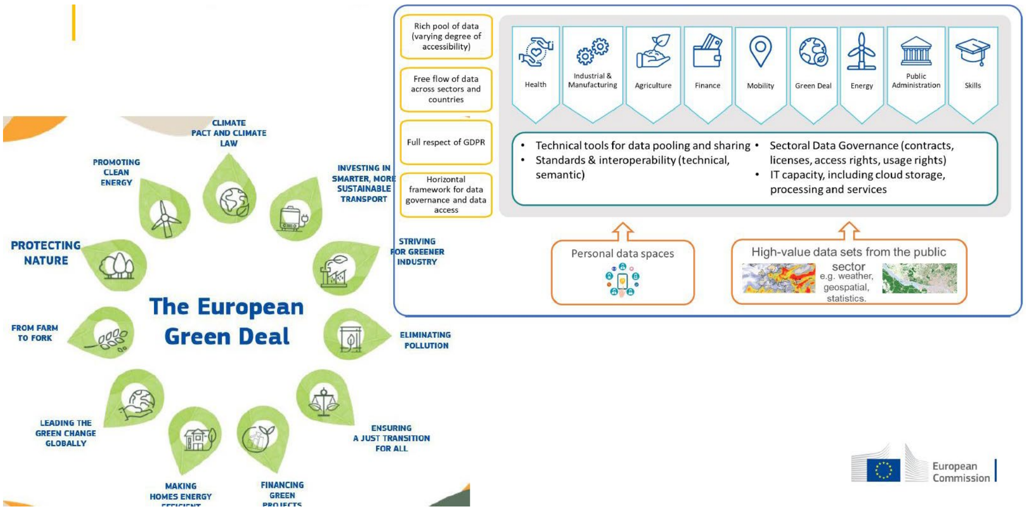

CLIMATE PACT AND CLIMATE LAW
PROMOTING CLEAN ENERGY
INVESTING IN SMARTER, MORE SUSTAINABLE TRANSPORT
PROTECTING NATURE
STRIVING FOR GREENER INDUSTRY
The European Green Deal
FROM FARM TO FORK
ELIMINATING POLLUTION
LEADING THE GREEN CHANGE GLOBALLY
ENSURING A JUST TRANSITION FOR ALL
MAKING HOMES ENERGY EFFICIENT
FINANCING GREEN PROJECTS
Rich pool of data (varying degree of accessibility)
Free flow of data across sectors and countries
Full respect of GDPR
Horizontal framework for data governance and data access
Health
Industrial & Manufacturing
Agriculture
Finance
Mobility
Green Deal
Energy
Public Administration
Skills
Technical tools for data pooling and sharing
Standards & interoperability (technical, semantic)
Sectoral Data Governance (contracts, licenses, access rights, usage rights)
IT capacity, including cloud storage, processing and services
Personal data spaces
High-value data sets from the public sector
e.g. weather, geospatial, statistics.
European Commission

FIGURE 3: EUROPEAN GREEN DEAL AND DATA SPACES[16]

The general SDI principle 'search-find-bind' for geospatial data can be maintained but needs to be revised. Geospatial data models have often become even more complex in responding to new user requirements; collecting, updating and maintaining the data might be facilitated by new technologies. However, these developments are challenging for the end user, but data provision can be streamlined, supporting simple, easy-to-handle and flexible data delivery within the wider web ecosystem–via the use of APIs, e.g. by following OGC Web API guidelines[17]. Figure 4 drafts a possible architecture for INSPIRE 2.0, but yet is not complete. Further IT components, like a Geospatial Rights Management (GRM) tier, need to be discussed and introduced. GRM can open the infrastructure for 'closed' communities because access control and privacy tracking could be embedded in the SDI.

[16] https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/452667656/[PRES1]_MIG11_Policy_info.pdf

[17] OGC API principles: https://github.com/opengeospatial/OGC-Web-API-Guidelines

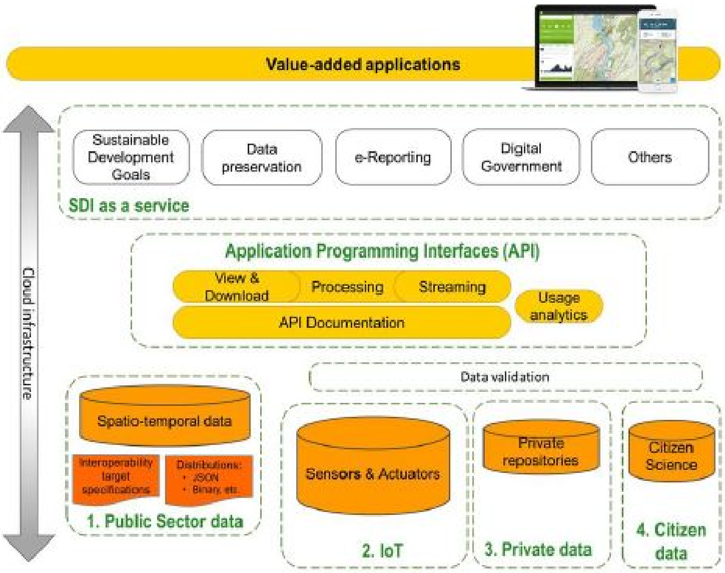


FIGURE 4: POSSIBLE ARCHITECTURE FOR INSPIRE 2.0 [18], ARTICLE IN THE INTERNATIONAL JOURNAL OF GEO-INFORMATION (ISPRS)

A general harmonisation of schema has already been done via INSPIRE, but a more semantics-based approach is still due. Geospatial data will be integrated by the use of common as well as domain-centric semantics. The data spaces are designed for specific issues and use cases, e.g. environmental reporting structures. Geospatial data will be made interoperable according to the requirements set in the respective data space – comprising the provision of ontologies and vocabularies for each data space if necessary. In order to achieve this, simple access solutions, sophisticated IT architectures and transparent methods for data integration within the data spaces are needed. The current deficits of INSPIRE with respect to semantics become obvious with the current landscape of harmonized versus non-harmonized 'as-is' datasets, which need to be considered in data integration. The issues associated require further evaluation and discussion, and cannot be covered in this document.

This is also the goal of the International Data Spaces Association (IDSA); with its IDS Architecture – as part of the strategy of the European Commission on the Strategic Value Chain of the Industrial Internet of Things (IoT) as well as of the strategy of the Digitizing European Industry (DEI) – IDSA is setting a standard for exchanging data on a trustworthy and self-regulated basis, a standard for data sovereignty[19].

[18] https://www.mdpi.com/2220-9964/9/3/176

[19] https://internationaldataspaces.org/

4 National Examples

To analyse different data integration methods and to provide further insights, as a first step in 2020, national examples in the European region were collected among the Working Group members. The aim of this exercise was to identify the most relevant methods in use among statistical and geospatial agencies. These national examples have been observed according to the following structure: method, additional requirements, implemented technology, involved agencies, used data and problems.
The WG has evaluated the methods and the embedded processes of each example –looking at specific strength or weakness– and has elaborated conclusive remarks for a long-term maintenance of data integration method in terms of organisational impact, data set maintenance and resource dependencies. Finally, the WG has addressed for each of those examples, main messages highlighting potential, requirements and characteristics.

These observations and analyses have been compiled in a table accessible at https://un-ggim-europe.org/wp-content/uploads/2021/07/Assessment-of-dataintegration-examples_-table_v20210714.pdf.

Based on these observations, a limited set of methods – highlighting the most relevant aspects on data integration – have been identified and explained, with a reference to the national examples.

5 Data Integration Methods

The chapter describes the methods most frequently used in the agencies for integrating statistical and geospatial data and taken from the national examples. The focus will be on different approaches to geocoding, the emerging developments of generic spatial join-operations, the usage of 'Linked Data'[20] and 'Ontologies'.

These methods, highlighting the most relevant aspects on (statistical-geospatial) data integration should be considered in the future development of national geospatial data infrastructures and of future geospatial knowledge infrastructures in general.

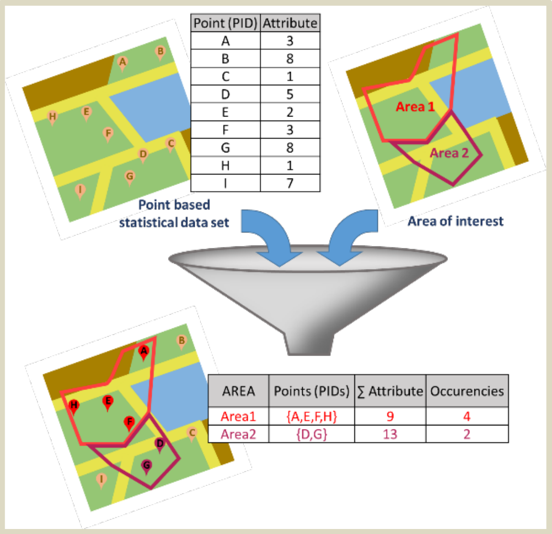
5.1 Point-based System

In the entire process of the integration of statistical and geospatial information, which is described in the Global Statistical Geospatial Framework (GSGF) document[21] and through the application of its five principles (see Annexe I), methods on data integration are of crucial importance.
One basic solution provided by the GSGF is called 'point-based' system. GSGF introduces it as base for enhanced statistical geospatial interoperability and data integration. The rationale of a point-based system is that it enables geocoding as basis of data integration based on location like point coordinates. Interaction of statistical and geospatial data with the point-based system can be seen on two levels:

[20] This paper focuses on the technical and not on the political aspect of Linked Data

[21] https://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf

the point-based geometries for data integration, and the expansion to the record unit level with Persistent Unique Identifiers (See Figure 5).



Point (PID)	Attribute
A	3
B	8
C	1
D	5
E	2
F	3
G	8
H	1
I	7

AREA	Points (PIDs)	∑ Attribute	Occurencies
Area1	{A,E,F,H}	9	4
Area2	{D,G}	13	2

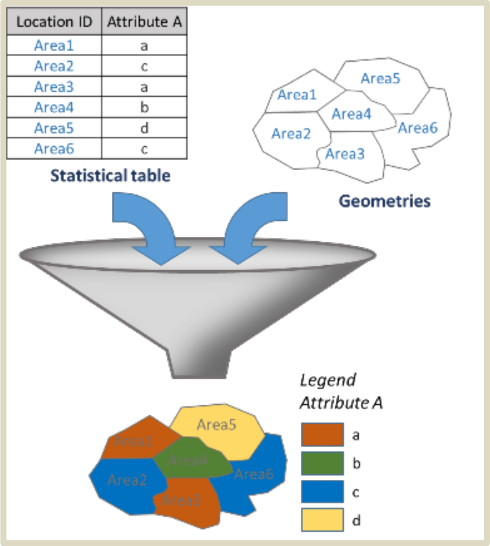
FIGURE 5: CONCEPT OF THE POINT-BASED SYSTEM

The point-based system is used in many European countries for various reasons with different advantages like the flexibility in terms of visualisation of geospatial data:

- Spain uses it to overcome a lack of data harmonization. The method allows the NSO to extract the cadastre data that are affecting the NSO without any change of data model in each institution.
- In Austria, spatial quality in the transport network graph is improved by an automated process for generating address points from cadastral parcels.
- In Germany, accessibility statistics at the family level (dwellings) were generated by rasterizing accessibility zones to the reference grid using this method.
- In Switzerland, this method makes it possible to geocode an important part of the statistical data by combining the geocoded register of buildings and dwellings, the links between the registers (population, businesses and establishments), their identifiers and a historical list of municipalities, as well as to generate aggregates for each historical state of the municipalities.
- In Finland, a pilot project on spatial indexing and geocoding has been created to integrate and aggregate data of record units with common persistent unique identifiers from different sources into territorial subdivisions and to further modify these spatial aggregates by performing spatial linkage even without GIS.
- In Poland, a pilot project created an internal Linked Open Data (LOD) implementation that links three types of data: statistical data (small sample), geospatial data (for spatial reference of statistical data) and metadata (for datasets).

5.2 Area-based Data Integration

Statistical tables as well as attribute data are usually labelled with area identifiers, which are also indirect references to location and geospatial data (see Figure 6). A standard method to merge geospatial portrayal of different area divisions with table-based attribute data is described by the OGC Standard Table Joining Service (TJS). TJS as such is a single-case integration method using common unique identifiers as links with the data to be integrated and may need a tender process between data providing organizations beyond open data.



Location ID	Attribute A
Area1	a
Area2	c
Area3	a
Area4	b
Area5	d
Area6	c

FIGURE 6: CONCEPT OF THE AREA-BASED DATA INTEGRATION

A more flexible and versatile solution is possible by publishing geographies of areal classifications and their relations as a data structure, which is feasible in Resource Description Framework (RDF) format (linked data) and possible also with data transformations from traditional databases (DBMS2RDF). This view on area-based data integration is to compile a structure for describing relations of areal classifications: one set of geographic regions and their nested relationship to other sets of (larger) regions, also called ‘allocation tables’ or regional relations.

Some examples illustrate the use of the area-based data integration for their application:

- In Finland, the IGALOD pilot provides geographies transformed into RDF and statistical classifications in the XKOS ontology, both of which are disseminated through SPARQL endpoints using a federated query; further development is needed to parameterize the input data model and the RDF model (see Figure 7).

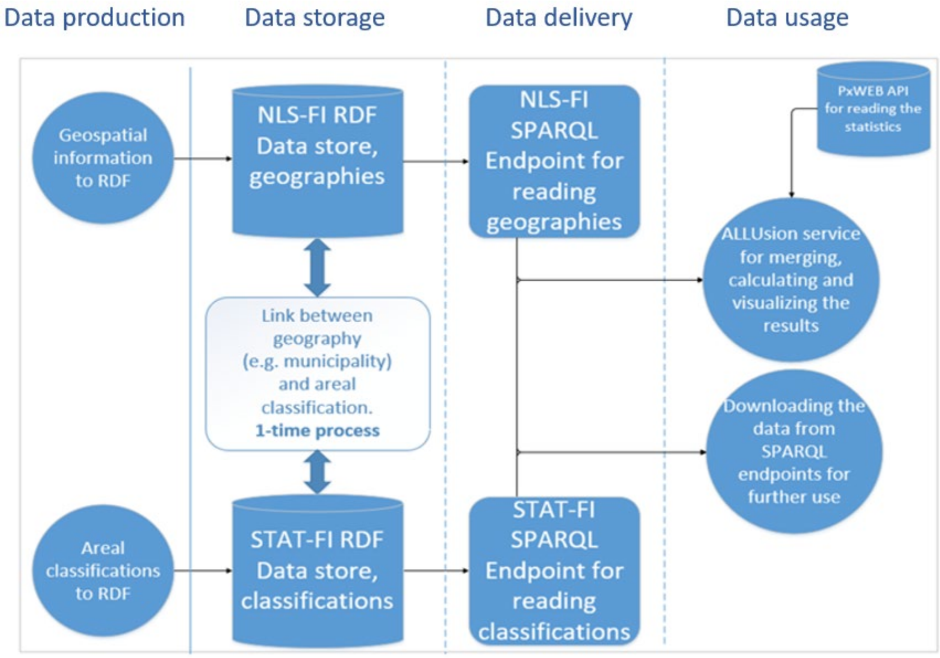

Data production
Data storage
Data delivery
Data usage
Geospatial information to RDF
NLS-FI RDF Data store, geographies
NLS-FI SPARQL Endpoint for reading geographies
PxWEB API for reading the statistics
ALLUsion service for merging, calculating and visualizing the results
Link between geography (e.g. municipality) and areal classification. 1-time process
Downloading the data from SPARQL endpoints for further use
Areal classifications to RDF
STAT-FI RDF Data store, classifications
STAT-FI SPARQL Endpoint for reading classifications

FIGURE 7: THE FINNISH IGALOD-SOLUTION IS USING FEDERATED SPARQL QUERIES TO ENABLE THE SPARQL-ENDPOINTS OF THE BOTH ORGANIZATIONS TO PROVIDE THE COMBINED DATA IN RDF FORMAT (STATISTICS FINLAND, NATIONAL LAND SURVEY)

- In Poland, a pilot study by Statistics Poland converts CSV files into RDF graphs and common statistical ontologies developed for Linked Open Data (LOD) as a source of georeferenced, machine-readable statistics.
- In Switzerland, the LINDAS project aims to develop the linked data service in all domains and at all levels of public administration and to create a service that can easily transform and visualize data into linked data.
- In the Netherlands, publishing Linked Data and SPARQL endpoints as interfaces to data models makes it possible to use spatial key registers and statistical data in federation: many data and topic-oriented use cases become accessible. An evolution to Knowledge Graph with APIs and facet browser Graphical User Interface (GUI) for specific analysis and 'Linked Data Maps' thus become possible.

5.3 Spatial Join Operation

Geospatial referencing and data integration necessitate common unique identifiers as geocodes. In addition, spatial aggregation can be performed also by location. This standard spatial join operation in a GIS follows the point-based system. A spatial join is a GIS operation that affixes data from one feature layer's attribute table to another from a spatial perspective.

Furthermore, a spatial join operation for the integration of statistical and geospatial data is possible with grid data. Therefore, geographical space is also defined as grid. It is important that this geospatial/statistical grid is congruent throughout the area that has to be analysed. Different definitions of grids may lead to incompatible analyses. For this reason, statistics in Europe have defined a statistical European grid system within INSPIRE[22].

The importance of the grid geocoding can be shown in the following examples:

- In Germany, accessibility statistics on family level (dwellings with location) were generated by rasterizing accessibility zones to the reference grid for aggregating accessibility.
- In Austria, similar work was performed in Austria for accessibility to education, retail, health, security and leisure (see Figure 8).

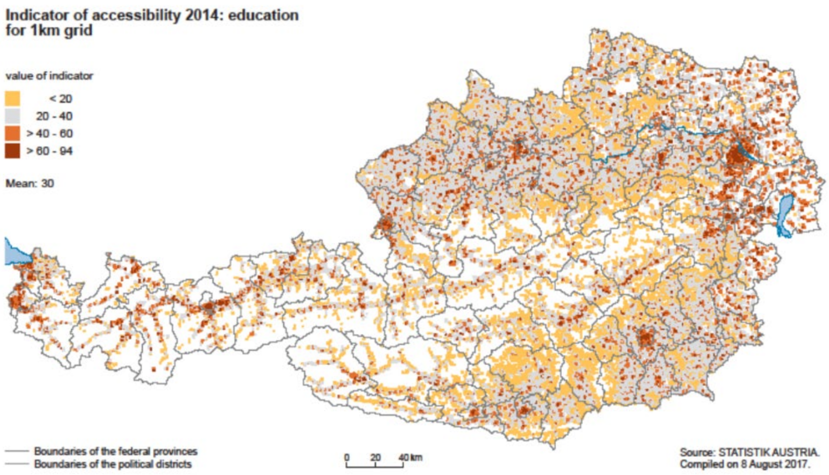

Indicator of accessibility 2014: education
for 1km grid
value of indicator
< 20
20 - 40
> 40 - 60
> 60 - 94
Mean: 30
Boundaries of the federal provinces
Boundaries of the political districts
0 20 40 km
Source: STATISTIK AUSTRIA.
Compiled on 8 August 2017.

FIGURE 8: ACCESSIBILITY GRID FOR EDUCATION GRANT: 08143.2015.001-2015.712

[22] https://inspire.ec.europa.eu/id/document/tg/gg; https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids

5.4 Persistent Unique Identifiers (PID)

A common characteristic and prerequisite for every implementation of data integration, especially the 'point-based' system, is deploying common persistent unique identifiers (PIDs) for statistical data on record unit level as means of geocoding by providing (indirect) spatial reference. Furthermore, PIDs promote management of semantics since PIDs identify the same data objects in different data repositories and vocabulary collections (see Figure 9).

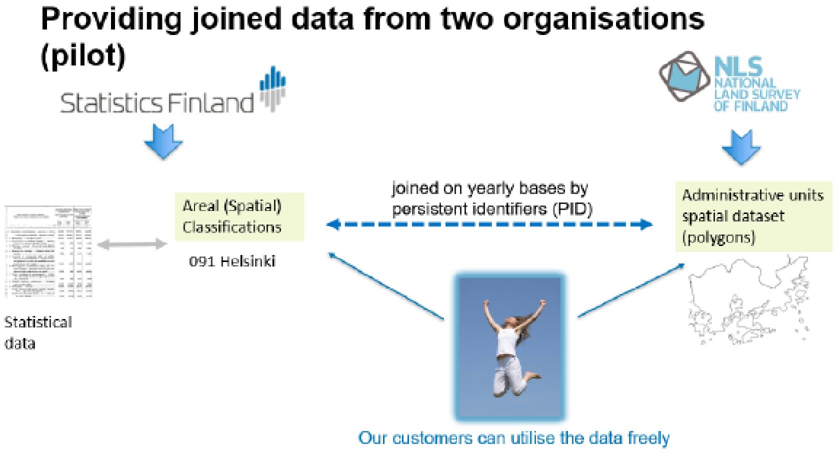

Providing joined data from two organisations (pilot)
Statistics Finland
NLS NATIONAL LAND SURVEY OF FINLAND
Areal (Spatial) Classifications
091 Helsinki
Statistical data
joined on yearly bases by persistent identifiers (PID)
Administrative units spatial dataset (polygons)
Our customers can utilise the data freely

FIGURE 9: DATA INTEGRATION EXAMPLE FROM FINLAND

In many countries, PIDs have been assigned to spatial database objects, especially on European level according to INSPIRE implementing rules on the INSPIRE spatial objects. To promote data linking beyond INSPIRE data and on national level and between different data providers, national guidelines on assigning PIDs and setting up a PID service with a PID (URI) resolver are needed. One example is published in Finland as a national recommendation[23] to the public administration. It contains annexes, which show and guide how to deploy and use PIDs. In addition, EU/ISA programme has published a profound study[24] on best practices on the publication of Uniform Resource Identifiers (URI) sets, both in terms of format and of their design rules and management.

It is often necessary to link key register data of SDI such as cadastral parcel, address, buildings and dwelling data in order to produce meaningful statistical information. Linking is performed explicitly through identifiers – e.g. NUTS code– or implicitly through geospatial relations – e.g. a building is within a commune. For example, dwellings and population registers will be linked to support policy questions like how many people live in energy efficient buildings. For the use case of routing – e.g. car navigation– the road identifier may be included in the address register. Such linking is activated and further enhanced with shared PIDs that enables geospatial re-aggregations e.g. spatial joins, particularly useful on record unit level, and can be expanded to different time versions by e.g. including time extensions in PID management or even use case defined versions of the PID.

[23] https://www.suomidigi.fi/en/jhs-193-unique-identifiers-geographic-information

[24] Study on persistent URIs: https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf

5.5 Ontologies

Ontologies are tools for managing semantics. As ontologies are built on a specific purpose and for thematic domains – there is no common ontology – an applicable solution is often a combination of custom ontologies that describe data lakes of closely related use cases and phenomena, as well as European data spaces and High Value Datasets (HVD). The current statistical/thematic classifications[25] and respective vocabularies[26] in the European languages could serve as a base for a common semantic grounding for ontologies because statistics are cross-sectoral and regarded as influential knowledge base (see Figure 10).

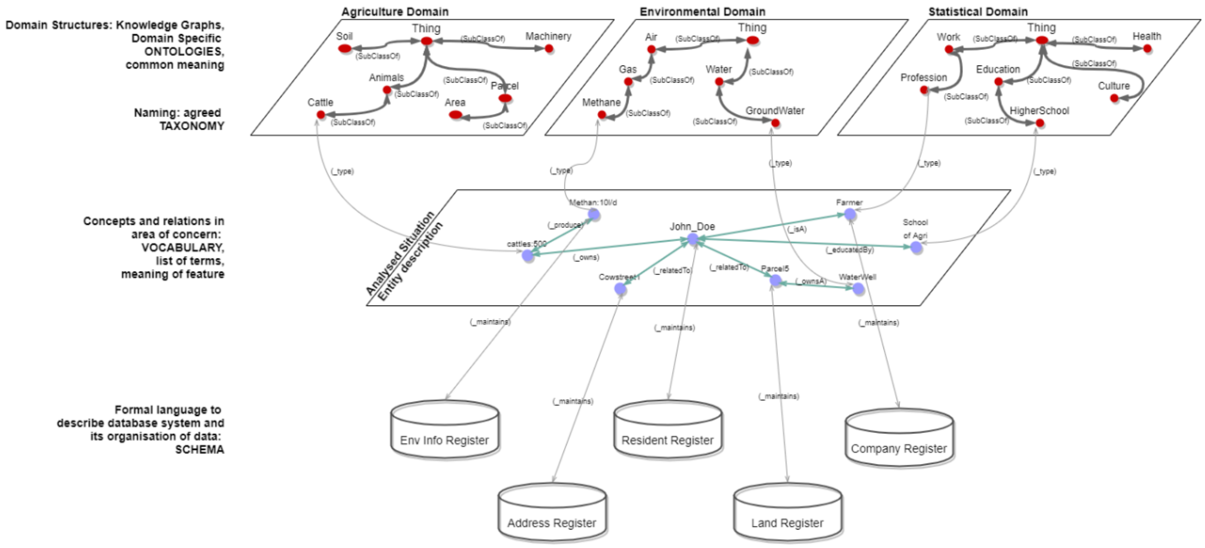


FIGURE 10: THE ROLE OF ONTOLOGIES, TAXONOMIES, VOCABULARIES AND DATABASE SCHEMAS IN A GRAPH-BASED KNOWLEDGE NETWORKS (SOURCE BY AUTHORS).

In addition to managing semantics via ontologies that describe how objects are related, a taxonomy classifies things and concepts. Taxonomy is understood as a kind of classification from a theoretical viewpoint. The taxonomy (classification) depends on the thematic domain, which creates the classification. Semantic overlaps between different taxonomies may occur. Geospatial fundamental data have the power to fuse different taxonomies on their georeferencing character – there is only one real-world object, although different descriptions and classes exist–, which is a key aspect of semantic data integration.

[25] https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC

[26] https://op.europa.eu/en/web/eu-vocabularies/authority-tables

5.6 Making Use of Resource Description Framework (RDF) for Linked Data and for Graph-based Linked Data

RDF is both format and data model to serve as technology for Linked Data. In addition to the method description on the 'area based data integration' (see chapter 5.2), the importance of 'RDF' and 'Linked Data' should be highlighted. Geospatial data, like geographies of areal classifications, and their data models can be transformed in RDF. The same applies to statistical data. Both models are then accessible on the web and can be integrated in a federated query through SPARQL, if this kind of endpoints are provided by each organization. Statistical data can be made accessible through SPARQL endpoints–e.g. in statistical Extended Knowledge Organization System (XKOS)[27] format–, including the geospatial relations and temporal versions and providing a continuous common semantic structure. Hence, linked data is most applicable for data integration from distributed and heterogeneous data sources provided with appropriate ontologies.

Another Data Integration convergence, the Graph-based Linked Data is gradually becoming a mature technology for statistical-geospatial data integration; the data are originally stored in graph structures / databases and not in relational tables (see Figure 10).

RDF for Linked Data represents an exchange interface and not a storage structure. A storage structure for graphs is established in a graph database e.g. by labelled property graphs. This allows for surpassing differences in data models, user interfaces and distributed data sources. It is providing unrivalled potential for data providers to enrich their information and knowledge supply as well as make their data capture more efficient e.g. through Knowledge Graphs and, step by step, evolving Knowledge Graphs Infrastructure[28].

In a number of NSOs (e.g. UK, PL, FR, CH, ES, BG, IE, IT) statistical data are already exposed in RDF format[29], while in the NMCAs the process is lagging. At the European level, the European Data Portal provides SPARQL/RDF endpoints to many European datasets.

In regard to the European spatial data infrastructure INSPIRE, the JRC has procured and launched the development of two pilot projects[30]. The main goal is to facilitate cross-sector interoperability and to help reuse the investments in the common SDI INSPIRE, which should also include available Linked Data interfaces and accessible Open Data portals.

While rapid development is underway, current standards are not fully taking advantage of linked information technologies. Different standards may not be compatible, and, for example, standards governing the definition and use of different coordinate systems and degrees of generalization are still

[27] https://ddialliance.org/Specification/RDF/XKOS

[28] Advancing Role of Geospatial Knowledge Infrastructure in World Economy, Society and Environment, page 20, https://geospatialmedia.net/pdf/GKI-Discussion-Document-Ver1.0.pdf

[29] https://ec.europa.eu/eurostat/cros/content/digicom-final-event-digicom-results-download_en

[30] https://joinup.ec.europa.eu/collection/are3na/news/using-inspire-geospatial-data

largely lacking. In the future, best practices on how to provide spatial datasets, associated metadata, and spatial relations on the web should be defined. This will be achieved by introducing a semantic tier based on knowledge graph networks in the information architecture. The main goal of the semantic tier is to create unambiguity between the different ontologies and their expression of meaning. An architecture and technology like this make it possible to link geospatial data in a versatile way to other PID referenced datasets and to search for and access spatial data on the web. Regarding spatial information, the work focuses especially on the development of vocabularies and ontologies. In addition, the spatial search capabilities of RDF database solutions are key areas for development from a spatial data perspective, where an important vehicle is Shapes Constraint Language (SHACL)[31,] a World Wide Web Consortium (W3C) specification for validating graph-based data against a set of conditions like filtering properties (geospatial and semantic attributes).

6 Data Integration Process

Defining concepts is not enough, and data integration is a long-term process. This chapter will highlight several interacting key processes that shall be considered so to run sustainable data integration process and which are, among others: the design of the surveys and data collections both for statistics and geographies, the common understanding of the semantics, the harmonisation of rules used for a coordinated maintenance of PIDs and temporal dimension, the use of common geographies, the simplification of the data models, and the automation of the data integration process in combination with the easy access and usage of the integrated data through standardized interfaces.
In addition, we have to notify that processes in matter of the reliability of web data sources and of the security of data transmission also have to be considered.

The process of data integration begins with the design of the surveys and data collections. The design of the statistics production should include the process of geocoding, which, in most cases does not start ab initio, but will be carried out by linking objects to already geocoded data such as buildings, addresses, roads, GPS coordinates, etc.[32]

This requires the use of shared PIDs and a common understanding of the semantics used for all datasets. Nevertheless, geographies often change over time, for example the NUTS regions of Europe. Therefore, it is also important to define how the PID evolves over time. Changes having an impact on those identifiers, these are documented and understood in the same way by all stakeholders involved, and their temporality is consistent; for example, if an object without geographic reference is geocoded by its link to a geographic feature, this geographic feature must exist at the time as the object is geocoded, and there must also be a process in place to ensure the integrity (or removal) of the link if the geographic feature is changed or deleted.

[31] https://www.w3.org/TR/shacl/

[32] Upcoming UNECE publication: GSBPM - Geospatial View 1.0

It means that each PID must be characterized by its temporal dimension. It is therefore mandatory to link geographies to a timestamp, and to ensure that PID from the geographies are also characterized by a temporal dimension.
The tools and interfaces available today make it easy to link (point-based) statistical data to geographical areas (grids, administrative areas, spatial statistical units, etc.). Therefore, the system must be ideally constructed in such a way that each state of the statistical time series can be represented in each state of the geographical time series.

As defined in the GSGF[33], the use of common geographies is a prerequisite for data visualization, analysis and interpretation of statistical data. Geospatial representation is more important than ever for making impact with thematic information, e.g. like statistics, and visualizing dependencies and relations to other information on selected questions. Geometries of areal classifications are originally created and updated in permanent procedures via GIS systems. This is directly done by SDI data providers or the geometries are forwarded to national geospatial repositories, e.g. of NSOs or NMCAs. This results in several challenges with respect to data integration; changes in areal boundaries shall be updated, the statistical record unit data shall be aligned to updated boundaries, timely different versions of this microdata and areas shall be kept, and sometimes the current view needs to be compared to past versions to demonstrate the actual change processes.

In some words, we may say that:

- Different pathways and gradual progress are possible depending on the circumstances. Every step should serve the next ones, the reusability on that pathway and the avoidance of partial solutions.
- Determination of common unique identifiers is the first step for each key data integration process
- Linked data developed to a Knowledge graph can be regarded as a cross-domain data repository and can help to foster productivity.
- During any production process it is important to check the integrity of the PID and links, both spatial and temporal, between the different datasets. All stakeholders must be involved in the production process and their collaboration is a condition for ensuring the quality of data integration.

Data integration should be an automated process and make use of standardized interfaces, e.g. open application programming interfaces (API). Processes can be simplified by renewing the whole workflow

[33] The GSGF is mostly and naturally drafted from the point of view of the statistical domain, while nowadays many organizations provide statistical information, not only to the statistical agencies, but directly to the users and for the benefit of society. This has also been discovered and recognised in ongoing GEOSTAT4-project.

or individual sub tasks, where new and innovative technical methods can be reused in several statistical tasks or workflows. Simplifying data models with formalized semantic representations (e.g. as linked data) and formats is replacing many redundant or single subtasks in similar workflows. As internal processes and data repositories become more and more aligned, huge benefits at cross-sector interoperability are achieved.

The access of data and the usage of data integration, which should also be automated, makes use of standardized interfaces, e.g. open application programming interfaces (API). It must be recognized that OGC API standards, here especially OGC API processes, are being developed to define resource-centric APIs to enable and remarkably facilitate geospatial data delivery and integrations on the web in various samples, formats –vector, raster, and coverage data–, area definitions and with filtering options on attribute data[34].
Similar to the development of service interfaces, formats and encoding standards are further promoted by other OGC standards –e.g. geopackage.[35.]

Moreover, recognizing that reliability of web data sources and the security of data transmission are significant concerns, Linked Data methods also allow the dissemination of malicious information and the falsification of information for malicious purposes (Linked Data Spam). Such threats can be combated e.g. through encryption, access control and malicious data filtering, and by addressing threats in the development of technologies and services. Attention must also be paid to ensuring data protection, as sensitive and personal data can also be processed and provided using the same technologies. To this end, standards like Data-centric security (DCS) by OGC are being developed.

7 Recommendations

Illuminating the data integration barriers from a more technical point of view within this paper, a call for action in Europe should address the following recommendations.

The paper structures the recommendations according to: their direct impact on the linked data integration processes –so to be successful in a technical way– and their belonging to a broader framework that influences the data integration environment.

7.1 Recommendations targeting the data integration methods and linked data processes

Altogether the basic recommendations for successfully linking geospatial and statistical data are:

1) Define and implement valid Persistent Identifier (PID) across domains requiring that:

[34] https://ogcapi.ogc.org/

[35] https://www.geopackage.org/

a) clear responsibilities for PID maintenance, relationships and stewardship have to be defined;

b) reference geometries have to be published with clear temporal reference – period of validity– and respecting the 'one geometry-one PID' relation. Versioned data including historic information – e.g. life cycle– should be harmonized across time and across different domains;

c) the publication process should be managed by authoritative agencies –e.g. NMCA. National and European Guidelines need to be developed. The definition of clear responsibilities can be supported with a specific role concerning persistent identity management;

d) PID consistency across datasets – e.g. PID must be the same between addresses, administrative unit, and statistical unit– must be ensured.

e) PID must be deployed at point and record unit level data.

2) Agree upon common definitions and enforcement of fundamental geographies and linked data:

a) sustainable harmonised common vocabularies and registries need to be established (ontologies);

b) fundamental geospatial data need to be established on a long term, trustworthy and self-regulated basis and respect the time space continuum of data delivery, removing legal and organisational constrains and minimizing various data structures;

c) geospatial data has to be connected by using common as well as domain-centric semantics in the data spaces, the latter designed for specific issues and use cases, e.g. environmental reporting structures.

3) Make geospatial and statistical data interoperable, simple, and of good quality in order to enable trustful and successful data integration processes throughout European data spaces:

a) geospatial data has to be made interoperable using more simple data schema – not too complex geometries–, but at the same time safeguarding the integrity and richness of the source data, and with specifications according to the requirements set in the respective data space – comprising the provision of ontologies and vocabularies for each data space if necessary;

b) data quality requirements should be part of the specifications which solve or even avoid data integration issues like gaps in data contents, or heterogeneous geometries;

    c) thanks to simplification the geospatial data can be easily integrated into the European data spaces;

    d) geospatial data has to be made available to the users via simple, easy-to-handle, standardized and flexible interfaces to accomplish cross domain usability of datasets.

4) Develop and implement sustainable and automated data integration processes:

    a) data integration processes should be implemented in a way that they can manage data updates and a temporal dimension of geographies;

    b) the processes should be automatized as much as possible in order to allow the easy access and the usage of the integrated data through standardized interfaces.

5) Develop and implement standardized Open APIs for a smart geospatial data provision comprising service interface and API definitions, styles and processes including geospatial rights management.

7.2 Recommendations belonging to a broader framework

The following recommendations act upon the knowledge infrastructure and upon the data infrastructure, which basically provide the suitable conditions for facilitating the data integration processes.

1) Invest resources and capacity building into a cross-domain ‘Geospatial knowledge infrastructure (GKI)’ aiming at common data spaces:

    a) geospatial and statistical organisations have to invest resources and capacity buildings in order to build knowledge, skills and expertise for the next generation of standards, tools and infrastructures to store, process and integrate harmonised and commonly licenced geospatial and georeferenced statistical data;

    b) geospatial and statistical organisations have to establish overarching governance structures in order to manage the evolution of the geospatial knowledge infrastructure.

2) Modernize the current National Spatial Data Infrastructures towards GKIs:

    a) NSDIs – and the geospatial data–have to be simplified for the integration into the European data spaces;

    b) key registers of SDI need to be linked via PIDs and common ontologies ensuring consistent temporality and integrity of datasets;

c) modernisation of the technology stack[36] is needed to embrace automation, dynamicity, high performance, reliability, security, and real-time delivery of knowledge;

d) security for ensuring data protection has to be put in place, because some of the data that is processed can contain sensitive and personal information;

e) organisations have to establish Knowledge graphs[37] to structure and document the way to transit and relate between the different approaches and data sources – geography, statistic, ontology;

f) best practices should be defined for graph based (meta) data dissemination and their relations.

8 Future perspectives

The UN-GGIM report "Future Trends in geospatial information management: the five to ten year vision"[38] provides a consensus view of the developments and future direction for geospatial information management over the next decade – comprising the challenges for data integration methods. It has been stated that "The next five to ten years will see significant developments in the maturity and application of already well-established technologies across the geospatial industry." Technology and methods will change the way in which data is collected, managed and maintained.

This report of the UN-GGIM: Europe Working Group on Data Integration underlines that Linked Data should be "seen as the key enabler for data integration". The consideration of Linked Data within national, regional or international Spatial Data Infrastructures (SDIs) will help for better and more efficient discovery, access, exploration and use of geospatial data through the Internet. Linked Data is able to establish a next generation of SDIs in the future. A wider concept towards geospatially enabled knowledge infrastructures is part of the future SDI. Experiences from Member States which already have implemented and adopted these concepts – like Finland, Switzerland, United Kingdom – as well as the European Commission have to be observed and good practice examples to be promoted and disseminated.

This report of the UN-GGIM: Europe Working Group on Data Integration is therefore a good starting point to further elaborate the perspectives for the data integration as being an essential part of a geospatially enabled knowledge infrastructure.

[36] Technology stack: stands for all parts (components) of the IT environment that are needed to establish a usable information system.

[37] The knowledge graph represents a collection of interlinked descriptions of entities – objects, events or concepts. Knowledge graphs put data in context via linking and semantic metadata and this way provide a framework for data integration, unification, analytics and sharing. https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

[38] http://ggim.un.org/meetings/GGIM-committee/10th-Session/documents/Future_Trends_Report_THIRD_EDITION_digital_accessible.pdf

Glossary

Abbreviation	Definition
4IR	Fourth Industrial Revolution
API	Application Programming Interface
CSV	Comma-separated values
DBMS	Data Base Management System
DCS	Data-centric security
DEI	Digitizing European Industry
EC	European Commission
EU	European Union
GEOSTAT	ESSnet project series since 2010, actually GEOSTAT-4
GIS	Geospatial Information System
GPS	Global Positioning System
GRM	Geospatial Rights Management
GSGF	Global Statistical Geospatial Framework
HVD	High Value Datasets
IDSA	International Data Spaces Association
INSPIRE	Infrastructure for Spatial Information in Europe
IoT	Internet of Things
JRC	Joint Research Centre (European Commission's science and knowledge service)
KGI	Knowledge Graphs Infrastructure
LOD	Linked Open Data
NMCA	National Mapping and Cadastral Agencies
NSO	National Statistical Office
NUTS	fr. Nomenclature des unités territoriales statistiques (Statistical Regions of the EU)
OGC	Open Geospatial Consortium
OSM	Open Street Map
PID	Persistent Unique Identifiers
RDF	Resource Description Framework
SDG	UN sustainable development goals
SDI	Spatial Data Infrastructures
SHACL	Shapes Constraint Language
SPARQL	Protocol And RDF Query Language
TJS	Table Joining Service
UN EG-ISGI	UN Expert Group on the Integration of Statistical and Geospatial Information
UN-GGIM	United Nations Global Geospatial Information Management
UN-GGIM: Europe	United Nations Global Geospatial Information Management Europe
URI	Unique Resource Identifiers
VGI	Volunteered Geographic Information

Abbreviation	Definition
W3C	World Wide Web Consortium
XKOS	Statistical Extended Knowledge Organization System

Annex I

The Global Statistical Geospatial Framework (GSGF)

To facilitate a consistent production and integration approach for geo-statistical information, the UN EG ISGI[39], has developed a Global Statistical Geospatial Framework (GSGF)[40]. Through the application of its five Principles (see Figure 11) and supporting key elements, the GSGF permits the production of harmonised and standardised geospatially enabled statistical data.

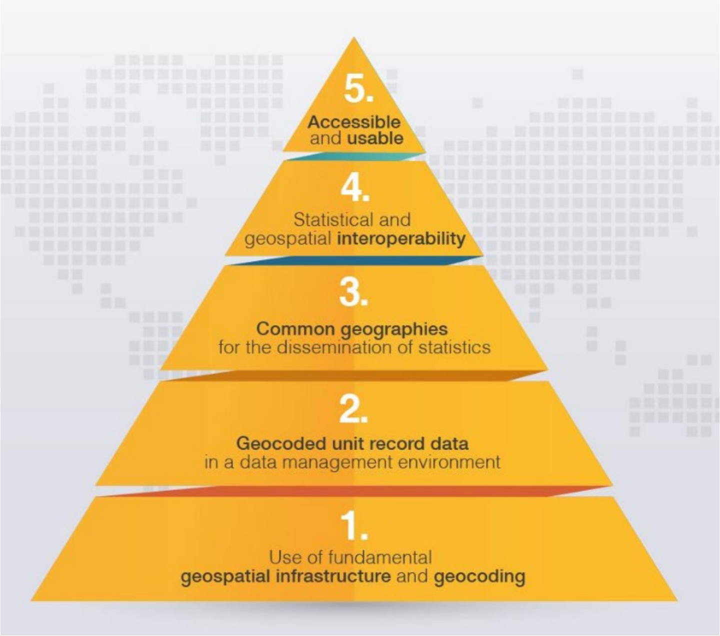

5.
Accessible and usable
4.
Statistical and geospatial interoperability
3.
Common geographies for the dissemination of statistics
2.
Geocoded unit record data in a data management environment
1.
Use of fundamental geospatial infrastructure and geocoding

Figure 11: Five principles of the Global Geospatial Framework[1]

The resulting data can then be integrated with statistical, geospatial, and other information to inform and facilitate data-driven decision making to support local and national development priorities and global agendas, such as the 2020 round of population and housing census and the 2030 Agenda for Sustainable Development.

At its simplest, the GSGF facilitates the integration of statistical and geospatial information, using location as a point of integration. In other words, data that relates to the same place can be related with each other. Importantly, the finer the location attribution on the data, the greater the accuracy that can be ascribed to that relationship. This integration allows data from the statistical and geospatial communities, as well as other information domains, to be brought together to understand our world and inform decisions made across our societies and nations.

[39] http://ggim.un.org/UNGGIM-expert-group/

[40] The GSGF was endorsed by the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) in August 2019, and the United Nations Statistical Commission in March 2020.