TOWARDS THE INTEGRATION OF AUTHORITATIVE AND OPENSTREETMAP GEOSPATIAL DATASETS IN SUPPORT OF THE EUROPEAN STRATEGY FOR DATA

A. Sarretta[a, *], M. Minghini[b]

[a] National Research Council, Research Institute for Geo-hydrological Protection, Padua, Italy - alessandro.sarretta@irpi.cnr.it
[b] European Commission, Joint Research Centre (JRC), Ispra, Italy - marco.minghini@ec.europa.eu

Commission IV, WG IV/4

KEY WORDS: Citizen-generated data, Europe, Interoperability, National Mapping Agencies, OpenStreetMap, Spatial Data Infrastructures

ABSTRACT:

Digital transformation is at core of Europe's future and the importance of data is well highlighted by the recently published European strategy for data, which envisions the establishment of so-called European data spaces enabling seamless data flows across actors and sectors to ultimately boost the economy and generate innovation. Integrating datasets produced by multiple actors, including citizen-generated data, is a key objective of the strategy. This study focuses on OpenStreetMap (OSM), the most popular crowdsourced geographic information project, and is the first step towards an exploration of pros and cons of integrating its open-licensed data with authoritative geospatial datasets from European National Mapping Agencies. In contrast to previous work, which has only tested data integration at the local or regional level, an experiment was presented to integrate the national address dataset published by the National Land Survey (NLS) of Finland with the corresponding dataset from OSM. The process included the analysis of the two datasets, a mapping between their data models and a set of processing steps—performed using the open source QGIS software—to transform and finally combine their content. The resulting dataset confirms that, while addresses from the NLS are in general more complete across Finland, in some areas OSM addresses provide a higher detail and more up-to-date information to usefully complement the authoritative one. Whilst the analysis confirms that an integration between OSM and authoritative geospatial datasets is technically and semantically feasible, future work is needed to evaluate enablers and barriers that also exist at the legal and organisational level.

1. INTRODUCTION

The digital transformation of the economy and society is at the very core of the European Commission's priorities for the period 2019-2024, centred around the twin need for a greener and more digital Europe (European Commission, 2019). This is also proven by the Recovery and Resilience Facility, recently established in response to the COVID-19 pandemic, which prescribes that at least 20% of the €672.5 billion provided to European Union Member States in loans and grants have to be used for the digital transformation (European Commission, 2021). Clearly, no digital transformation can happen without data and, reflecting this, the European strategy for data (European Commission, 2020a) envisions Europe's digital future through the establishment of a European single market for data ensuring the free flow of data, including personal and non personal, across actors and sectors, to stimulate data-driven innovation and create value for the economy and society. The vision is to establish a common European data space based on domain-specific data spaces in strategic sectors such as environment, agriculture, industry, health and transportation. To achieve this goal, an ambitious set of legislative instruments to be released by 2024 will address a number of data-related issues such as availability, interoperability, quality, governance, cybersecurity, skills and literacy as well as the overarching data infrastructures. The European strategy for data acknowledges the importance of all kinds of data, being them produced by the public sector, the private sector, academia or citizens. Hence, making it possible to combine and integrate data from different sources—by solving all the issues mentioned above—acquires primary importance for the successful establishment of data spaces.

This paper addresses the topic of integrating data produced from the public sector and from citizens, with a focus on the geospatial domain and within a European dimension in mind. In the European strategy for data, data contributed by citizens—a phenomenon referred to as 'data altruism'—play a central role and shall happen in full compliance with the General Data Protection Regulation (European Parliament and Council, 2016). The potential of citizen-generated data to improve policy making has been already widely recognised by the European Commission, e.g. in the fields of citizen science (European Commission, 2020b) and, more specific to the geospatial domain, Spatial Data Infrastructures, where citizen-generated data contributes to their evolution into modern geospatial data ecosystems (Kotsev et al., 2020).

This study explicitly focuses on citizen-generated data from OpenStreetMap (OSM), the most well-known and successful crowdsourced geographic information project. Started in 2004 and currently (June 2021) counting more than 1.6 million unique contributors (https://wiki.openstreetmap.org/wiki/Stats), OSM consists of a global database of geospatial vector features available under the open access Open Database License (ODbL). Thanks to the freedom of use ensured by the license, as well as its richness and level of detail, the OSM database is currently used by a variety of actors including governments, private companies and nonprofit organisations (Mooney and Minghini, 2017). The problem of integrating OSM with other datasets, mainly authoritative datasets produced by governmental National Mapping Agencies (NMAs)—which is discussed in this paper—has been addressed since the very early OSM literature in close connection with research on OSM quality; notable examples include Haklay (2010), Girres and Touya (2010) and Neis et al. (2012). Several experiments were carried out on specific features (roads, buildings, land use areas, etc.) and using OSM and authoritative data from many regions in the world. However, those experiences still appear iso-

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-159-2021 | © Author(s) 2021. CC BY 4.0 License.

lated as they mostly describe specific use cases, are only tested on small (local or regional) areas, are bounded to particular authoritative datasets and often rely on data model-dependent procedures, which are hard, if not impossible, to generalise and replicate.

With this background, this work aims to be the first step towards a broad assessment of the enablers and barriers of integrating authoritative datasets from European NMAs with datasets from OSM. The overall purpose is to provide a preliminary set of recommendations on interoperability matters, not only semantic but also technical, organisational and legal, to ultimately support the establishment of European data spaces. To achieve this, the study proposes an experiment based on Free and Open Source Software for Geospatial (FOSS4G) to test the integration of country-wide address datasets from a European NMA and the OSM project, discussing the outcomes and identifying lessons learnt and general pros/cons of data integration mainly from the technical perspective. To the authors' knowledge, this is the first time the integration between OSM and authoritative datasets at the national level is addressed in literature. Evaluating the quality of OSM clearly remains a key and preliminary step to such integration, but is outside the scope of the study; an extensive review on how OSM quality has been measured so far is available in literature (Senaratne et al., 2017).

The remainder of the paper is structured as follows. After an analysis of the state of the art on the integration between authoritative and OSM datasets provided in Section 2, Section 3 describes the experiment of integration between the authoritative dataset of national Finnish addresses and its OSM counterpart, adopting FOSS4G technology. Drawing from the results of the experiment, Section 4 closes the paper by discussing implications of, and providing recommendations on, the integration of citizen-generated data (and OSM in particular) for the successful establishment of data spaces.

2. BACKGROUND ON INTEGRATION BETWEEN AUTHORITATIVE AND OPENSTREETMAP DATA

Being a citizen-driven project, OSM has been studied—and sometimes questioned—since its very beginning in relation to the quality of its data. This aspect was first addressed by some early studies, e.g. Haklay (2010) and Girres and Touya (2010), who described and measured various quality parameters on OSM data through in-depth assessments, e.g. attribute, semantic, positional and temporal accuracy, logical consistency, completeness, lineage, purpose and usage. Quality assessment methods are of course not only relevant to the case of OSM but, more generally, for all types of Volunteered Geographic Information (VGI) (Senaratne et al., 2017). Many other studies investigated those different quality elements, focusing on the semantic (Vandecasteele and Devillers, 2013) and positional (Cipeluch et al., 2010; Helbich et al., 2012) aspects, completeness (Koukoletsos et al., 2012), interoperability (Minghini et al., 2019) or, more frequently, on a combination of them, e.g. Fan et al. (2014).

Most of the available studies on OSM quality adopted an extrinsic approach, i.e. they compared OSM data with reference datasets produced by National Mapping Agencies (NMAs) or local, national or international authoritative bodies that are considered as the ground truth. Fernandes et al. (2020) provided a bibliometric review of 37 studies on the integration between VGI and authoritative data, even if only 14 of them use OSM as the main source for VGI. Among them Du et al. (2012), Abdolmajidi et al. (2014), Fan et al. (2016) and Brovelli et al. (2017) developed and tested methodologies to evaluate the quality of OSM data by comparing it against their authoritative counterparts, using the road network as a use case applied at the local level (city or town) in different places around Europe (UK, Sweden, Germany and Italy, respectively). Instead of comparing OSM with authoritative datasets, other studies such as Barron et al. (2014), Minghini and Frassinelli (2019) and Madubedube et al. (2021) assessed OSM quality through intrinsic approaches, i.e. by only looking at the history of the OSM data itself (e.g. the frequency of update or the total number and nature of contributors editing the same objects).

Nevertheless, just a few authors have focused their efforts on combining authoritative and/or OSM data together to produce integrated datasets. This conflation process involves different tasks, which can include updating, change detection, enhancement and integration of spatial data (Wiemann and Bernard, 2010). Pourabdollah et al. (2013) compared OSM and the British Ordnance Survey's Vector Map District data on road network. Differently from many other authors, who focused their attention on geometrical accuracy and completeness, they focused on semantic information, conflating road names and reference codes with the main result to enrich the OSM dataset with authoritative information. The potential contribution of OSM data to the increase of mapped features of the authoritative road network in Brazil was the goal of Silva et al. (2021): their analysis confirmed that OSM is a promising source of information in areas with missing or outdated map data. Zhou et al. (2015) presented instead an extensive method used to dynamically integrate OSM data from the neighbouring states Vietnam and Pakistan into a common data model. Other studies focused on the semantic enrichment of authoritative datasets by extracting information from specific OSM tags related to building usage (residential/non-residential), e.g. Kunze and Hecht (2015). Similarly, Fonte et al. (2017a) developed an automated, FOSS4G-based application to convert OSM into land use/cover maps having the same nomenclature of authoritative products. This allowed not only to compare the OSM-derived products against the authoritative ones, but also to enrich the latter through the production of integrated datasets (Fonte et al., 2017b). However, the most frequent and structured case of integration between OSM and authoritative datasets to date is represented by so-called OSM imports, or bulk imports (https://wiki.openstreetmap.org/wiki/Import). These consist of uploading external datasets, produced e.g. by governments or other institutions and having a license compatible with the ODbL, into the OSM database. Imports are tricky operations and shall be performed based on specific guidelines issued by the OSM community (https://wiki.openstreetmap.org/wiki/Import/Guidelines); an updated list of OSM imports performed so far is maintained at https://wiki.openstreetmap.org/wiki/Import/Catalogue.

3. INTEGRATION EXPERIMENT: DATA SOURCES

The selection of the authoritative dataset to be integrated with OSM plays an important role in the phases of analysis and harmonisation of data models, the transformation process and its possible reuse for other areas or use cases. The dataset selected in this work to test the integration approach between authoritative and OSM data is about addresses. In addition to being usually modelled as points with a reasonably simple data model, addresses represent reference datasets for a multitude of applications. They are not only a core dataset produced and maintained by governments at all levels, but also one of the most important datasets within the OSM ecosystem, considering e.g. the wealth of OSM-based routing or emergency applications (Mooney and Minghini, 2017). Furthermore, addresses represent a typical case where the process of updating the authoritative dataset is traditionally expensive and not frequent and might thus highly benefit from an integration with OSM.

While the study maintains a European perspective for the general issue of integrating authoritative and citizen-generated datasets, as mentioned in Section 1 the scale of the experiment was limited to a national geographical area for both computational and semantic reasons. This is in contrast with all the studies mentioned in the literature review presented in Section 2, which have been always limited to more restricted (local or regional) areas. Given the focus on address data, we identified Finland as a useful and practical example because of: i) the easy access to the authoritative address dataset, and ii) the wide coverage of OSM addresses. The two address datasets used in the experiment are described in the following Sections 3.1 and 3.2 together with their main characteristics and modes of access.

3.1 OpenStreetMap

OSM data is organised using a simple conceptual data model combining a geometric component with a semantic component (Ramm and Topf, 2011). The geometric component can be described using three types: nodes, ways and relations. Nodes are characterised by a latitude and a longitude and represent standalone point features such as points of interest, trees, street signals and benches; ways are an ordered list of up to 2000 nodes representing both linear features (e.g. roads and rivers) and areal features or polygons (e.g. buildings and land cover areas); relations are data structures used for both modelling linear and areal features with more than 2000 nodes (e.g. lakes) or describing a relationship between two or more geometry types (nodes, ways and/or other relations), e.g. transportation networks. The semantic component consists of one or more attributes, named tags and each formed by a key-value pair.

Information on how addresses are modelled in OSM is available at https://wiki.openstreetmap.org/wiki/Addresses. The keys of all the tags used to identify addresses share the common `addr:` prefix (https://wiki.openstreetmap.org/wiki/Key:addr). The keys associated with address information used in this experiment are described in Table 1. Other address-related keys available in OSM are `addr:unit`, `addr:postcode`, `addr:suburb`, `addr:state`, `addr:province`, `addr:floor`, `addr:place`, etc.

OSM tag	Description
`addr:country`	country code of the address
`addr:city`	name of the city of the address
`addr:street`	name of the street of the address
`addr:housenumber`	building number of the address

Table 1: Address-related OSM keys used in this experiment.

From the geometrical perspective, there is no single way to model OSM addresses. The `addr:` keys can be associated to single nodes outside, inside or on the perimeter of a building footprint; or they can be directly associated to the ways representing building polygons. Such different mapping practices are usually agreed upon by local, regional or national OSM communities and may also follow rules issued by national registry/statistical services. In the case of OSM addresses in Finland, all the abovementioned approaches are used and there seems to be no specific internal rule agreed upon by the community on how to perform mapping on this object category. In addition to that, address information in OSM can be also added to points of interest like shops, museums, offices, etc., sometimes duplicating addresses already available in other objects.

Extracting data from the OSM database can be performed in different ways, depending on the user needs. The most popular ones include: i) the use of Application Programming Interfaces (APIs), e.g. the OSM API (https://wiki.openstreetmap.org/wiki/API) and the Overpass API (https://wiki.openstreetmap.org/wiki/Overpass_API); ii) the download of predefined OSM extracts, e.g. provided by GeoFabrik (https://download.geofabrik.de) or the Humanitarian OpenStreetMap Team (https://export.hotosm.org/en/v3); and iii) the Planet OSM, a weekly-updated copy of the whole OSM database (https://planet.openstreetmap.org). For the purpose of this work, OSM addresses were extracted from the Planet OSM, downloaded on 7 June 2021 in the binary Protocol Buffer File (PBF).

3.2 National Land Survey of Finland

The National Land Survey (NLS) of Finland is the Finnish NMA (https://www.maanmittauslaitos.fi/en) . As such, it is the Finnish governmental provider of and responsible for the national geospatial information. The NLS has recently started to provide access to its geospatial datasets through the newly established OGC API - Features standard (https://ogcapi.ogc.org/features), which provides an easy and developer-friendly way to both expose and consume geospatial vector features on the web. The OGC API - Features service endpoint for addresses (https://beta-paikkatieto.maanmittauslaitos.fi/inspire-addresses/features/v1) is currently (June 2021) in beta version, is open and free of charge and does not require registration, but both the service and related materials are only available for testing. The service follows the recently developed INSPIRE (European Parliament and Council, 2007) Good Practice for the provision of INSPIRE download services based on OGC API - Features (https://github.com/INSPIRE-MIF/gp-ogc-api-features) and the address data exposed by the API are compliant with the INSPIRE Addresses data specifications (INSPIRE Thematic Working Group Addresses, 2014) and the INSPIRE UML-to-GeoJSON encoding rule (https://github.com/INSPIRE-MIF/2017.2). The NLS address dataset is available in the WGS84 geographic coordinate reference system according to the OGC API - Features standard and the GeoJSON specification Internet Engineering Task Force (2016). The draft data model is published at https://tietomallit.suomi.fi/model/ostieto and will be refined during 2021. Addresses are modelled as point features; among all the available attributes (which also include INSPIRE-specific information on e.g. identification and temporal context), those specifically related to addresses are listed in Table 2. The NLS address dataset is available under the open access CC BY 4.0 license (Creative Commons, 2021a).

NLS attribute	Description
`component_ThoroughfareName_name_fin`	name of the street of the address in Finnish
`component_ThoroughfareName_name_swe`	name of the street of the address in Swedish
`component_ThoroughfareName_name_sme`	name of the street of the address in Sami
`locator_designator_addressNumber`	building number of the address
`component_AdminUnitName_4`	code of the city of the address
`component_AdminUnitName_1`	country name of the address

Table 2: Address-related NLS attributes.

Figure 1 shows a portion of the two address datasets from OSM and the NLS in the area of Helsinki. The figure confirms that, in some cases, OSM address tags are associated to the building

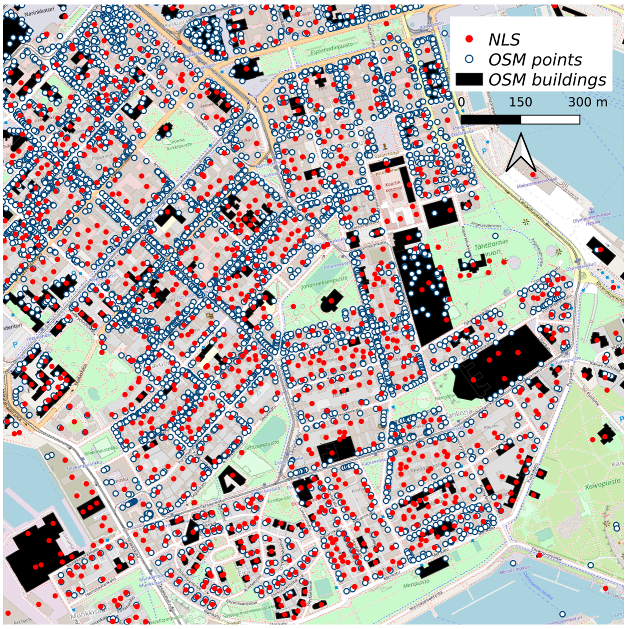

Figure 1: Example of distribution of address data in an area of Helsinki, Finland: OSM addresses associated to nodes (white points) and ways (black polygons); NLS addressed (red points). Background map: © OpenStreetMap contributors.

polygons. Also, it is visually clear that OSM addresses in this area, as it usually happens in urban areas (see also Section 4.2), are more than NLS addresses.

4. INTEGRATION EXPERIMENT: APPROACH AND RESULTS

4.1 Integration process

This section describes the procedures implemented to pre-process the OSM and NLS address datasets, mainly to extract the relevant information described in Sections 3.1 and 3.2, and to integrate them into a single dataset. Given that the data model for address data is richer in INSPIRE (that the NLS dataset conforms to) than in OSM, we considered that the best approach for the integration of the two was to transform the INSPIRE-compliant NLS dataset against the OSM data model. This was a fully arbitrary choice; the opposite one, i.e. the transformation of the OSM dataset against the NLS/INSPIRE data model (corresponding to the use case of an NMA wishing to complement its dataset with information from OSM) would be possible as well. All the steps described in the following were applied as a sequence of processing algorithms within the Graphical Modeler of the open source QGIS software (https://qgis.org) and are publicly shared on an online repository (https://github.com/MarcoMinghini/INSPIRE-OSM) to maximise their re-use and improvement.

In the case of OSM, a number of steps were performed to extract the relevant information from the OSM Planet and make it available in a format suitable for integration with NLS data. The Osmium Tool (https://osmcode.org/osmium-tool) was used to filter the Planet OSM both geographically (on Finland) and semantically, the latter by only extracting objects with a non-null value for the `addr:housenumber` key. The resulting dataset, transformed in the GeoPackage format, included both points (OSM nodes) and polygons (OSM ways) for the reasons explained in Subsection 3.1. Polygons were converted to points using their centroids and then merged with the pointwise addresses in a unique point dataset.

Several OSM address objects did not include the key `addr:city` filled with a value. Thus, this information was retrieved from the Local Administrative Units (LAU) dataset, downloaded from the Eurostat GISCO website (https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/lau) and then processed (since it originally included names in different languages) to match the existing information in OSM. Other OSM addresses were instead lacking the street name (key `addr:street`) and, similarly to those without the building number, were excluded from the dataset. After this process, OSM objects having the same unique combination of values for the keys `addr:city`, `addr:street`, `addr:housenumber` and `addr:unit` were considered duplicated and were removed from the dataset. Some additional minor processing steps were performed on the OSM dataset, but they are only described in the code in order not to make reading more difficult.

To transform the NLS address dataset against the OSM data model, a mapping between the NLS/INSPIRE and the OSM attributes was first required. This is shown in Table 3.

INSPIRE/NLS attributes	OSM attributes
`locator_designator_-addressNumber`	`addr:housenumber`
`component_-ThoroughfareName_name_fin`	`addr:street`
`component_AdminUnitName_4`	`addr:city`
`component_AdminUnitName_1`	`addr:country`

Table 3: Mapping between attribute names of the INSPIRE/NLS and the OSM data models related to addresses used in this work.

The three attributes that, at a national level (i.e. inside the same country), uniquely identify an address are the city name, the street name and the address number. With regard to the address number, both the `locator_designator_addressNumber` attribute in the NLS dataset and the `addr:housenumber` attribute in the OSM dataset store it as a string including the number (plus additional elements such as letters, e.g. *12b*). To align the two values, a simple rename of the NLS attribute was sufficient. Instead, the name of the street is documented in 3 attributes in the NLS dataset: `component_ThoroughfareName_name_fin` (corresponding to the name in Finnish), `component_ThoroughfareName_name_swe` (corresponding to the name in Swedish) and, lastly, `component_ThoroughfareName_name_sme` (corresponding to the name in Sami). We selected the first (see Table 3) whenever available (i.e. 99% of the times) and the second otherwise. The third one (name in Sami) was never used as it did not appear in any object. In the case of the city name, the value of the NLS dataset attribute `component_AdminUnitName_4` is a number representing the code id of the LAU (instead of its name). The name was thus retrieved from the LAU dataset and then substituted to the city id. To complete the transformation, the NLS attribute `component_AdminUnitName_1` (indicating the country) was renamed `addr:housenumber` and its values, all equal to *Finland*, were simply substituted with the ISO 3166-1 alpha-2 two letter country code in upper case (*FI*) in accordance with the OSM rules. As a last step, all duplicated addresses (i.e. addresses having exactly the same city, street and housenumber), which were sometimes appearing within different buildings close to each other, were identified and removed. The pre-processed OSM and NLS address datasets were finally merged into a single, integrated dataset with the basic rule to keep the attribute values from the NLS dataset in all the cases where the values of the fields `addr:city`, `addr:street` and `addr:housenumber` were the same in the two datasets. Figure 2 summarises the pro-

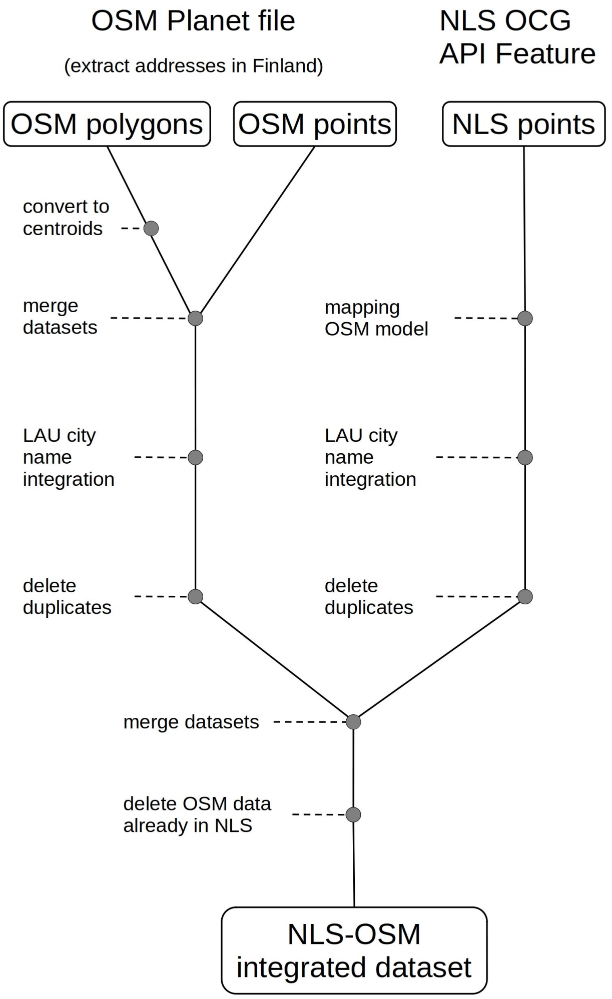

Figure 2: Graphical model of the processing performed to integrate the OSM and NLS address datasets in a single dataset.

cessing steps described above, which were implemented inside the QGIS Graphical Modeler.

4.2 Results

The two original datasets are collected and updated through very different procedures, thus it is not surprising that they also have large differences in the number of objects mapped and their distribution across the country. The NLS dataset, which was harmonised to the OSM data model, included around 3.3 million addresses, while the OSM dataset had just over 0.5 million (about 390000 polygons and 130000 points). The removal of duplicates brought the number of addresses down to 1.8 million for NLS and around 0.4 million for OSM.

The relative geographical distribution of the datasets is also very uneven. Considering the NLS address dataset as the reference one, Figure 3 shows that OSM data is in general much less complete, with a high variety of patterns. The 10x10 km EEA reference grid (https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2) was used to aggregate data, count the number of OSM and NLS addresses included in each cell and compute their percentage ratio. Approximately 63% of the cells where there is at least one address in the NLS dataset do not contain any address in the OSM dataset (white squares in Figure 3); the percentage ratio is less than 10% for about 24% of the cells and between 10% and 50% for another 7% of the cells. In slightly more

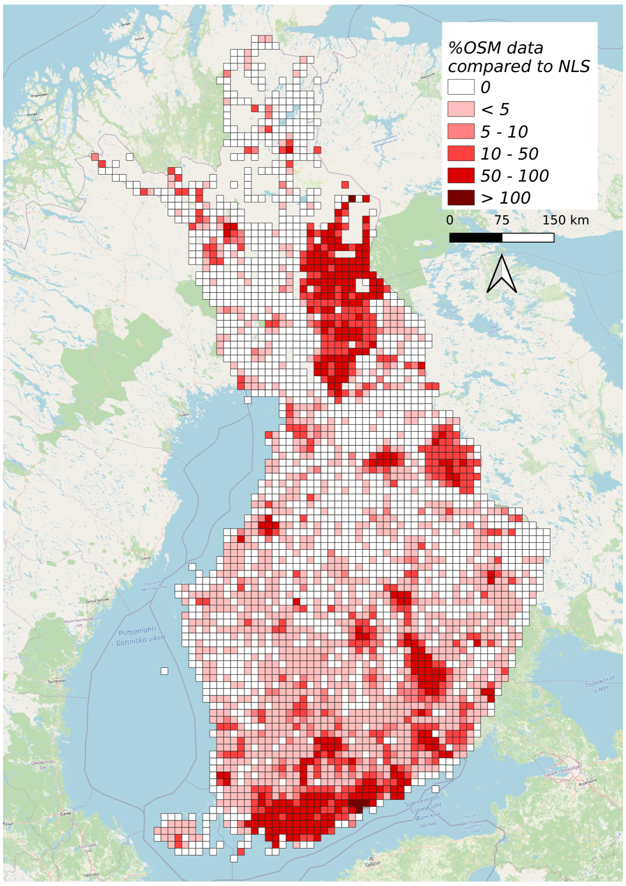

Figure 3: Percentage ratio between the number of OSM and NLS addresses, computed on the 10x10 km EEA Reference Grid. Background map: © OpenStreetMap contributors.

than 6% of the cells, the percentage ratio grows between 50% and 100% and only a few cells include more addresses in OSM than in the NLS dataset (percentage ratio higher than 100%).

Some of the most densely populated areas (based on the 2019 population figures included in the LAU dataset) are among the administrative areas that are most complete in OSM: 4 among the 6 most populated Finnish cities (Helsinki, Espoo, Vantaa and Turku) have average percentage ratios ranging between 75% and 97%. This confirms some typical findings from the literature, showing that areas with higher population densities (i.e. urban areas) tend to be those where most OSM mappers add and update information as they either live of visit such areas, see e.g. Zielstra and Zipf (2010), Dorn et al. (2015) and Brovelli et al. (2016). In addition to that, in some of those cities extensive OSM imports from authoritative sources have been performed in the past, thus highly increasing the number of addresses. As an example, an import of buildings that also included address information was performed starting in 2014 in the whole Helsinki region (https://wiki.openstreetmap.org/wiki/Helsinki_region_building_import).

The final, integrated address dataset includes around 1.92 million address points, with 96% of them being only present in the original NLS dataset and approximately 81000 of them only present in OSM. It should be clarified that this high number includes several cases where the name of streets or cities is mispelled (or spelled differently) in OSM with respect to the NLS dataset, which may highlight weaknesses in the OSM dataset rather than gaps in the one from NLS. However, there are also cases where OSM actually includes more detailed or up-to-date information and thus

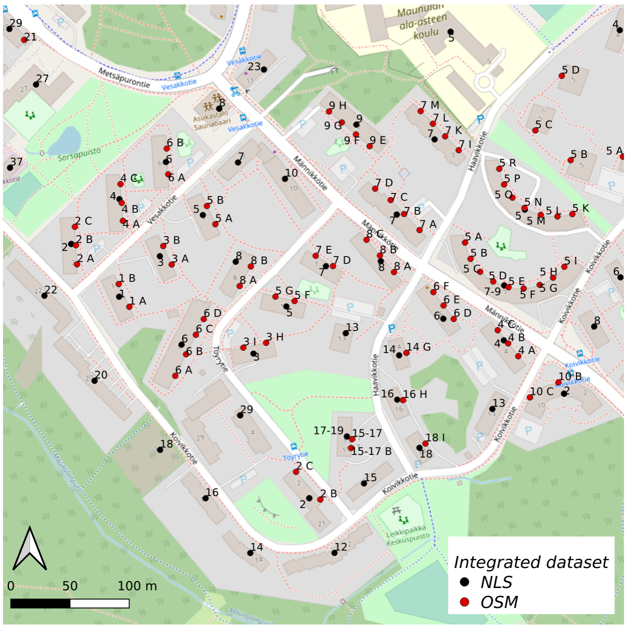

Figure 4: Integrated address dataset in an area in Helsinki showing the origin of each address point: OSM dataset (red), NLS dataset (black). Background map: © OpenStreetMap contributors.

improves the authoritative NLS dataset. As an example, Figure 4 shows an area in Helsinki where addresses in the NLS dataset, each associated to a single building, correspond to multiple addresses in the OSM dataset, where the building numbers are complemented by letters (*A*, *B*, *C*, etc.) and have a more specific position, most probably in correspondence of the single building entrances.

In addition to the QGIS Graphical Modeler workflow, the online repository at https://github.com/MarcoMinghini/INSPIRE-OSM also includes a sample of the final, integrated address dataset limited to the city of Helsinki for demonstration and testing purposes.

5. DISCUSSION AND CONCLUSIONS

Despite being simple, the experiment presented in this paper is useful enough to understand the complexity inherent to the process of integrating datasets which differ in nature and content. As such, the lessons learnt are a good first step to formulate helpful recommendations for the successful establishment of the data spaces envisioned in the European strategy for data (European Commission, 2020a).

First and foremost, any data integration process should be carefully prepared. This means that the datasets to be integrated shall be well-known in terms of their creation/update process, geometric representation, encoding, semantic content and quality (measured, in principle, through all the parameters that are important for the integration). If quality information is not available a priori, then a preliminary quality assessment becomes the first key step. This work deliberately assumed that the quality of the OSM address dataset across Finland was such that a comparison and integration with the NLS address dataset was actually possible without a dedicated, in-depth quality assessment. This was mainly justified by the very local nature of OSM, which allows to assume that the positional accuracy of OSM addresses is sufficiently high. In contrast, the possible low degrees of OSM address completeness (i.e. lack of addresses in some parts of the country) and semantic accuracy (i.e. wrong or missing address information) are directly taken into account in the integration process.

From the purely technical perspective, which was the focus of this work, a number of conclusions can be drawn. Results show that the integration between the OSM and NLS address datasets could improve both datasets, since the integrated dataset was achieved by 'taking the best' from both the initial ones. In general, results show that, while authoritative data have a more homogeneous coverage and higher positional accuracy, OSM has typically an uneven spatial coverage but holds the potential to include more updated or detailed information that authoritative datasets can only achieve, if ever possible, in a much longer time. This means that, in general, both the NMA and OSM communities might benefit from such integrations for improving their data. Ideally, such integration processes could be automated and executed on a regular basis to achieve increasingly more updated and higher-quality datasets.

As mentioned earlier, one of the main contributions of this work is that the integration between OSM and authoritative data happened at the national level, in contrast to previous work that was all focused on the regional or local scale (see Section 2). The experiment also showed that, although integration procedures involving OSM data are in general hard to generalise because of the peculiar nature and characteristics of the authoritative datasets involved (see again Section 2), the interoperability ensured by INSPIRE would allow the process to be seamlessly extended to other INSPIRE-compliant address datasets available across the EU.

From the software perspective, the experiment described proved that FOSS4G, and in particular QGIS and its Graphical Modeler, is a fully suitable Extract-Transform-Load (ETL) tool to perform the data processing involved in the integration (see Section 4). However, given the focus on nationwide datasets, it is worth mentioning that the process required a minimum computational capacity as it dealt with huge amounts (millions) of address features, which—if extended to all Europe—would need a proper infrastructure in place.

As mentioned in Section 1, this experiment is the first step within a broader research framework investigating enablers and barriers for the integration between authoritative and citizen-generated (in particular OSM) datasets in Europe. As such, it only focused on some interoperability aspects (technical and semantic) required for the integration, but it did not address other aspects such as the legal and organisational ones. Legal interoperability looks at dataset integration from the perspective of their licenses and terms of use. Whilst integration might be technically possible, the lack of license compatibility might indeed represent a serious obstacle to the actual use of the integrated datasets. This applies in both directions. To be integrated in OSM, a dataset shall have a license compatible with the ODbL: examples of such licenses include CC0 (Creative Commons, 2021b), while other licenses are either not compatible or (as in the case of NLS's CC BY 4.0) not compatible in the absence of an additional waiver for reasonable attribution and unrestricted distribution (https://wiki.openstreetmap.org/wiki/Import/ODbL_Compatibility). NMAs might face similar issues, since OSM's ODbL requires the release of the integrated dataset under the same ODbL license, which might be against existing national policies. In this regard, the recently published Open Data Directive (European Parliament and European Council, 2019) has pushed the publication of so called 'high-value datasets' (i.e. data-sets the re-use of which is associated with high economic and societal benefits) under open licenses, which should favour their integration with other data sources such as OSM. The final list of high-value datasets, together with the requirements for their provision (including the license), will be provided in a legal act foreseen for late 2021. In addition to legal interoperability, organisational interoperability both within and across organisations (including governments and OSM communities) will be key to make data integration a common, standardardised and policy-enabled process rather than an isolated and ad hoc exercise.

As a final note, readers should be aware that the definition of OSM as a citizen-generated database is increasingly challenged. Not only governments and other organisations have largely contributed to OSM through imports, but today more and more private companies using OSM for their business are heavily adding OSM data through their paid staff (Anderson et al., 2019). Hence, while still remaining a citizen-driven initiative, OSM has evolved into a broad and complex ecosystem with both the need to refine its governance and the potential to maintain and improve what is currently one of the most used global datasets worldwide.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the European Commission - Joint Research Centre (JRC) through contract number CT-EX2013D133588-101, entitled “Evaluation of Novel approaches for governing (location) data and technology. Combined use of public sector and citizen-generated data”.

DISCLAIMER

The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

References

Abdolmajidi, E., Will, J., Harrie, L. and Mansourian, A., 2014. Comparison of matching methods of user generated and authoritative geographic data. In: 17th ICA Workshop on Generalization and Multiple Representation, Vienna, p. 15.

Anderson, J., Sarkar, D. and Palen, L., 2019. Corporate Editors in the Evolving Landscape of OpenStreetMap. ISPRS International Journal of Geo-Information 8(5), pp. 232. https://www.mdpi.com/2220-9964/8/5/232.

Barron, C., Neis, P. and Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Transactions in GIS 18(6), pp. 877–895.

Brovelli, M. A., Minghini, M., Molinari, M. and Mooney, P., 2017. Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets. Transactions in GIS 21(2), pp. 191–206.

Brovelli, M. A., Minghini, M., Molinari, M. E. and Zamboni, G., 2016. Positional accuracy assessment of the OpenStreetMap buildings layer through automatic homologous pairs detection: The method and a case study.

Cipeluch, B., Jacob, R., Winstanley, A. and Mooney, P., 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In: Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences, p. 4.

Creative Commons, 2021a. Creative Commons Attribution 4.0 International (CC BY 4.0). https://creativecommons.org/licenses/by/4.0 (accessed on 23 June 2021).

Creative Commons, 2021b. Creative Commons 0 1.0 Universal (CC0 1.0) Public Domain Dedication. https://creativecommons.org/publicdomain/zero/1.0 (accessed on 24 June 2021).

Dorn, H., Törnros, T. and Zipf, A., 2015. Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. ISPRS International Journal of Geo-Information 4(3), pp. 1657–1671.

Du, H., Anand, S., Alechina, N., Morley, J., Hart, G., Leibovici, D., Jackson, M. and Ware, M., 2012. Geospatial Information Integration for Authoritative and Crowd Sourced Road Vector Data: Authoritative and Crowd Sourced Road Vector Data. Transactions in GIS 16(4), pp. 455–476.

European Commission, 2019. The European Commission's priorities for 2019-24. https://ec.europa.eu/info/strategy/priorities-2019-2024_en (accessed on 7 June 2021).

European Commission, 2020a. Commmission Staff Working Document: Best Practices in Citizen Science for Environmental Monitoring. https://ec.europa.eu/environment/legal/reporting/pdf/best_-practices_citizen_science_environmental_monitoring.pdf (accessed on 7 June 2021).

European Commission, 2020b. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European Strategy for Data. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020D C0066&from=EN (accessed on 7 June 2021).

European Commission, 2021. Public Sector Modernisation for EU Recovery and Resilience. https://ec.europa.eu/jrc/en/science-update/public-sector-modernisation-eu-recovery-and-resilience (accessed on 7 June 2021).

European Parliament and Council, 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002.

European Parliament and Council, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN (accessed on 7 June 2021).

European Parliament and European Council, 2019. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). Official Journal of the European Union L 172, pp. 56–83.

Fan, H., Yang, B., Zipf, A. and Rousell, A., 2016. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. International Journal of Geographical Information Science 30(4), pp. 748–764.

Fan, H., Zipf, A., Fu, Q. and Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science 28(4), pp. 700–719.

Fernandes, V. O., Elias, E. N. and Zipf, A., 2020. Integration of authoritative and Volunteered Geographic Information for updating urban mapping: challenges and potentials. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLIII-B4-2020, Copernicus GmbH, pp. 261–268. ISSN: 1682-1750.

Fonte, C. C., Minghini, M., Patriarca, J., Antoniou, V., See, L. and Skopeliti, A., 2017b. Generating Up-to-Date and Detailed Land Use and Land Cover Maps Using OpenStreetMap and GlobeLand30. ISPRS International Journal of Geo-Information 6(4), pp. 125.

Fonte, C. C., Patriarca, J. A., Minghini, M., Antoniou, V., See, L. and Brovelli, M. A., 2017a. Using OpenStreetMap to create land use and land cover maps: Development of an application. In: Volunteered Geographic Information and the Future of Geospatial Data, IGI Global, pp. 113–137.

Girres, J.-F. and Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. Transactions in GIS 14(4), pp. 435–459.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: Planning and design 37(4), pp. 682–703.

Helbich, M., Amelunxen, C., Neis, P. and Zipf, A., 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In: GI_Forum 2012: Geovizualisation, Society and Learning, pp. 24–33.

INSPIRE Thematic Working Group Addresses, 2014. D2.8.I.5 Data Specification on Addresses – Technical Guidelines. https://inspire.ec.europa.eu/id/document/tg/ad (accessed on 22 June 2021).

Internet Engineering Task Force, 2016. The GeoJSON Format. https://datatracker.ietf.org/doc/html/rfc7946 (accessed on 22 June 2021).

Kotsev, A., Minghini, M., Tomas, R., Cetl, V. and Lutz, M., 2020. From Spatial Data Infrastructures to Data Spaces—A technological perspective on the evolution of European SDIs. ISPRS International Journal of Geo-Information 9(3), pp. 176.

Koukoletsos, T., Haklay, M. and Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data: Assessing Data Completeness of VGI. Transactions in GIS 16(4), pp. 477–498.

Kunze, C. and Hecht, R., 2015. Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. Computers, Environment and Urban Systems 53, pp. 4–18.

Madubedube, A., Coetzee, S. and Rautenbach, V., 2021. A Contributor-Focused Intrinsic Quality Assessment of OpenStreetMap in Mozambique Using Unsupervised Machine Learning. ISPRS International Journal of Geo-Information 10(3), pp. 156.

Minghini, M. and Frassinelli, F., 2019. OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date? Open Geospatial Data, Software and Standards 4(1), pp. 9.

Minghini, M., Kotsev, A. and Lutz, M., 2019. Comparing INSPIRE abd OpenStreetMap data: how to make the most out of the two worlds. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLII-4-W14, Copernicus GmbH, pp. 167–174.

Mooney, P. and Minghini, M., 2017. A review of OpenStreetMap data. In: G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte and V. Antoniou (eds), Mapping and the Citizen Sensor, Ubiquity Press, pp. 37–59.

Neis, P., Zielstra, D. and Zipf, A., 2012. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007-2011. Future Internet 4(1), pp. 1–21.

Pourabdollah, A., Morley, J., Feldman, S. and Jackson, M., 2013. Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network. ISPRS International Journal of Geo-Information 2(3), pp. 704–728.

Ramm, F. and Topf, J., 2011. OpenStreetMap: using and enhancing the free map of the world. UIT Cambridge, Cambridge.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C. and Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science 31(1), pp. 139–167.

Silva, L. S. L., Camboim, S. P., Silva, L. S. L. and Camboim, S. P., 2021. Authoritative cartography in Brazil and collaborative mapping platforms: challenges and proposals for data integration. Boletim de Ciências Geodésicas.

Vandecasteele, A. and Devillers, R., 2013. Improving volunteered geographic data quality unsing semantic similarity measurements. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XL-2-W1, Copernicus GmbH, pp. 143–148. ISSN: 1682-1750.

Wiemann, S. and Bernard, L., 2010. Conflation Services within Spatial Data Infrastructures. In: Proceedings of AGILE 2010, Guimarães, Portugal, p. 8.

Zhou, X., Zeng, L., Jiang, Y., Zhou, K. and Zhao, Y., 2015. Dynamically Integrating OSM Data into a Borderland Database. ISPRS International Journal of Geo-Information 4(3), pp. 1707–1728.

Zielstra, D. and Zipf, A., 2010. A comparative study of proprietary geodata and Volunteered Geographic Information for Germany. In: 13th AGILE International Conference on Geographic Information Science, Vol. 2010.