Data (Re)conciliation: Audit and Harmonization Pipelines for Municipal Geospatial Data

Rodrigo João Bastos Mendes[1], Ertuğrul Doğruluk[2], Alexandre Valente Conceição Pereira de Sousa[3,4], and Pedro Correia Cravo Pimenta[3,4,5(✉)]

[1] Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal
up202308366@fc.up.pt
[2] Department of Software Engineering, Kahramanmaras Istiklal University, Kahramanmaras, Türkiye
ertugrul.dogruluk@istiklal.edu.tr
[3] Department of Computer Science, Universidade da Maia, Maia, Portugal
{asousa,ppimenta}@umaia.pt
[4] LIACC - Artificial Intelligence and Computer Science Laboratory, Porto, Portugal
[5] N2i - Núcleo de Investigação do Instituto Politécnico da Maia, Maia, Portugal

Abstract. Processing urban data frequently includes a significant effort to combine and reconcile data that is, by its nature, multi-source, fragmented, and updated frequently. This raises severe interoperability challenges and impacts data quality, mainly in its actuality, integrity and completeness. We also noticed that in Maia, as in many other municipalities, mismatches occur between the official geo-referenced data and the data available on crowd-sourced platforms such as OpenStreetMap or Overture. This paper presents a multi-source approach that, by design, aims to use all available relevant sources and combines the best from each. We developed a suite of automated pipelines that pull data from a variety of sources such as the authoritative city's GeoServer, GTFS feeds and official PDF documents and crowd-sourced systems such as OpenStreetMap and Overture. The framework leverages Python and DuckDB-over-S3 for in-memory Big Data processing, as well as unstructured PDF scraping. Our core contribution is a hybrid matching algorithm combining geospatial proximity analysis (Haversine formula) with domain-specific string similarity (Levenshtein-based token-set ratios) to classify cross-source relationships into: perfect matches, location conflicts, name conflicts, or missing entities. We applied this methodology across six urban domains processing over 1,500 spatial entities using real data from the Municipality of Maia, Portugal. Results validate the system's effectiveness, achieving a 91% alignment rate in operational transit networks, while exposing severe positional and semantic inconsistencies, including spatial deviations exceeding 500 m in global datasets and up to 692 m in unstructured alternative formats. By producing reproducible spreadsheets and interactive mapping tools, this research operationalizes data governance, empowering non-technical municipal technicians with "human-in-the-loop" validation capabilities, and proposes a roadmap toward a resilient, data-driven Urban Data Space.

Keywords: Geospatial Data · Data Reconciliation · Urban Data Space · Crowd-sourced Geographic Information · Data Quality

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 555–568, 2027.
https://doi.org/10.1007/978-3-032-30521-3_35

1 Introduction

The paradigm of the Smart City is fundamentally predicated on the continuous flow, integration, and analysis of urban data. City data services and citizens need to know where the schools are, which bus stops are active, and where the nearest health clinic is located. Geographic Information Systems (GIS) make this possible, but there is a catch: although in most cases the data is accurate, sometimes discrepancies exist in designation, address or geographic coordinates, or in a combination of these. In our work with the Municipality of Maia, we found that official Spatial Data Infrastructures (SDIs) are struggling to keep up. While citizen-led projects like OpenStreetMap (OSM) [1] get updated by volunteers, the city's internal Web Feature Service (WFS, GeoServer [2]) might take some time to reflect changes, such as a moved bus stop or a change in the name of a School.

This is a clear example of the so-called "data fragmentation" challenge - we can argue that we found the official truth in city hall services, the public truth on OSM [1], and the corporate truth in datasets like Overture Maps [3]. On top of this, not all data sources expose exactly the same data fields - one dataset may include an address (but not Lat, Lon), and the other might have (Lat, Lon) but not the full name of the building/service, etc. Integration of GIS and non-GIS data is thus a well-known issue in the area of geographical statistics [4] - data collection and linking process relies on expert knowledge and domain-specific understanding to combine spatial and non-spatial dimensions, which increases processing time and human expertise costs [5]. Streamlining the data collection, processing, harmonization and enrichment procedures to deliver high-fidelity datasets readily usable for analytics is therefore paramount for both national statistical offices and local municipal administrations - and for the citizens - alike.

Within the context of the D4Maia (Data for Maia) initiative at the Câmara Municipal da Maia (Portugal), our goal was to build a system that acts as a "bridge" between these different data worlds, assessing how far can we go by automatically processing these different datasets, taking into consideration eventual disagreements and complementarities and reporting both matches and mismatches in a summary that could be as easily as possible handled by data curators in the city data services. We developed several Python pipelines that take the official city datasets and compare them with external complementary datasets, through the key steps described below:

1. accessing city's GeoServer data services [2] and "Guia de Recursos Sociais do Município da Maia" (pdf) [6];
2. accessing OpenStreetMap [1] data endpoint - Overpass.api [7];
3. using DuckDB [8] to extract local data (Maia) from Overture [3];
4. definition of a hybrid matcher, considering both physical distance (Haversine formula) and Levenshtein distance[9] on names (fuzzy matching);
5. application of previous steps to six different datasets as metro stations, bus stops, schools, social services, health services and hydrography elements.
6. finding and reporting of matches and mismatches in both interactive maps and spreadsheets;

The data processed covers six datasets in an overall count of 1,500 entities, and main results are i) interactive maps where geographical matches and mismatches can be easily identified, and ii) spreadsheets that non-technical staff can actually use as a guide to fix location (Lat, Lon) errors, name or designation or description mismatches.

The remainder of this paper is organized as follows: Section 2 reviews related work on spatial data integration; Section 3 details the heterogeneous data sources and ingestion strategies; Section 4 presents the overall architecture supporting the hybrid matching framework and Section 5 discusses the case studies and quantitative results across urban domains. Finally, Section 6 provides a discussion on the implications of our findings for data governance, and Section 7 presents a discussion and directions of future work.

2 Background and Related Work

The Quality Gap in Spatial Data. Merging map data is an old and well-known issue for GIS engineers. Historically, most of the focus was on matching simple shapes [10], and Neis et al. [11] showed that volunteer data on OSM is often quite accurate, sometimes more accurate than official government records in certain areas. We build on this idea, not as a comparison tool of 'right' vs. wrong, but as an audit tool to report the conflicts to be addressed later by qualified personnel.

Automation and the Human Element. De Paoli et al. [5] argue that semantic enrichment with geospatial data needs a "human-in-the-loop" - the algorithm should find the potential matches, but a human must be able to verify them easily. Hassani's team, building GISINTEGRATION to link GIS layers to standard databases, reported that this linking is hard because the formats and coordinates do not always match [4], thus reinforcing the need for expert judgment for further analysis.

New Global Standards. We are also seeing a shift toward the availability of global/large datasets. Overture Maps [3] is publishing satellite data and POIs through cloud-native Parquet files. Metro do Porto [12], a supra-municipal metropolitan transport operator serving seven municipalities in the Porto Metropolitan Area, publishes its GTFS through Porto Digital [13]. For the local context of Maia, the usage of these types of overall, large datasets is not straightforward; we need a mix of spatial formulas (like Haversine [14]) and text similarity tools (like Levenshtein [9]) to extract relevant data for our local context.

Record Linkage in Geography. To reconcile entities across datasets without shared unique identifiers, systems must rely on spatial and attribute similarity, and Chopde and Nichat [14] demonstrated the efficacy of this approach in landmarks reconciliation. However, spatial proximity alone might not be sufficient in dense urban environments and thus semantic similarity algorithms, such as Levenshtein distance [9], are required to disambiguate nearby entities (e.g., distinguishing a pharmacy from an adjacent hospital clinic).

This hybrid approach echoes the confidence-score-based candidate selection in SemTUI [5], where uncertain reconciliation results are surfaced for human review. Our research combines these approaches into a unified, domain-adaptable framework - the

key distinction is operational: SemTUI targets interactive, user-driven tabular enrichment, whereas our pipelines are designed for fully automated, periodic auditing of a municipal SDI.

3 Data Sources and Ingestion Strategies

Data sources were already identified as relevant and present a design-chosen diversity, viz: i) authoritative sources as Maia Municipality and Metro do Porto, ii) crowd-sourced as OpenStreetMap and iii) mixed, foundation-governed as Overture. The main challenges were dealing with the diversity of formats (pdf, GTFS, REST API, Parquet) and reducing all the coordinates to EPSG:4326 (WGS84). As Hassani et al. [4] noted, this kind of standardization requires coordinate transformation and semantic reconciliation before any comparison can occur.

3.1 Official Municipal Data (The Baseline)

Maia uses a GeoServer [2] to share its official geospatial data. We used their WFS endpoints to download GeoJSON files for schools, medical centers, and social services / social service buildings. A first issue we noticed was that the raw WFS layers sometimes lack proper administrative names. To fix this, we joined the spatial points with national datasets like the "GesEdu" database [15] to recover the correct names.

3.2 Crowdsourced Geographic Databases (The Community View)

To extract community-maintained data, we used the Overpass API [7] to OSM to run tag-specific queries, restricting the spatial scope to within Maia's boundary polygon[1]. For example, we used tags like `highway=bus_stop` for our bus stops script and `medical amenity` or `pharmacy` or `clinic` or `hospital` tags for our healthcare POIs audit. Data is returned in json format.

3.3 Overture Maps (The Big Data Challenge)

Overture Maps [3] releases planet-wide data in Parquet format hosted on Amazon S3. Downloading the entire dataset locally is computationally prohibitive and thus Overture provides a DuckDB interface (through `httpfs` protocol). This allows direct SQL-formatted queries execution over the remote S3 buckets. A first spatial scope should be specified through a bounding box (bbox.xmin, bbox.xmax, etc.) in the WHERE clause to pull the POIs inside that box (Listing 1).

```
SELECT names.primary AS Name, ST_AsWKB(geometry) AS geometry
FROM read_parquet('s3://overturemaps-us-west-2/release/2024-05-16-
beta/theme=places/+')
WHERE bbox.xmin BETWEEN -8.68 AND -8.49
  AND bbox.ymin BETWEEN 41.20 AND 41.28;
```

[1] D4Maia endpoint: https://baze.cm-maia.pt/BaZe/api/api4gj.php?nome=LimConcVM.

Data pulled with this query, scoped for a rectangle-boxed limits, have to be further filtered for the specific polygon of Maia Municipality limits so to keep only the points within the Maia Municipality.

3.4 Operational Transit Data

We used Porto Digital [13] services to pull GTFS files. Most of the time the system responded flawlessly but eventually we came across two error situations: i) the GTFS files served by Porto Digital were in a zero-length condition, and ii) the server returned a 403 error. For these two situations, we implemented a fallback mechanism that catches these errors and loads a previously healthy GTFS file saved in Google Drive.

3.5 Alternative Information Dissemination Formats (PDF, XLSX)

For social services, we scraped the "Guia de Recursos Sociais" (a PDF document, [6]). The whole process consists of three steps, viz: i) using `pdfplumber` [16] to extract text blocks; ii) using regular expressions to isolate service and organization names and addresses; and iii) using Nominatim [17] REST API to geocode the POI. In the case of the list of schools, available through GesEdu [15], the platform provides a sophisticated menu for choosing the desired schools based on several attributes, and finally saves the schools list in xlsx format.

4 Methodology: The Hybrid Matching Framework

4.1 Computational Environment

The pipelines were developed and operated in google colab [18]. At the time of the project, google colab environment provided execution environments with standard cloud CPUs, 12GB of RAM and Python 3.10. We relied on widely used Python modules not included in the Python 3.10 standard libraries as `geopandas` [19] and `shapely` [20] for geospatial data manipulation and `FuzzyWuzzy` [21] for Levenshtein distance calculations. Finally, DuckDB [8] (version 0.9+) was used to query data directly from Overture's AWS S3 buckets.

4.2 Pipeline Architecture

Overall, our reconciliation system currently consists of six specific Python notebooks[2]. The architecture follows a linear Directed Acyclic Graph (DAG), illustrated in Figure 1, moving from i) ingestion, ii) hybrid reconciliation and iii) output generation.

The matching algorithm combines two factors in a two-step process:

[2] https://github.com/RodrigoMendes04/PI2025-DataHarmonization

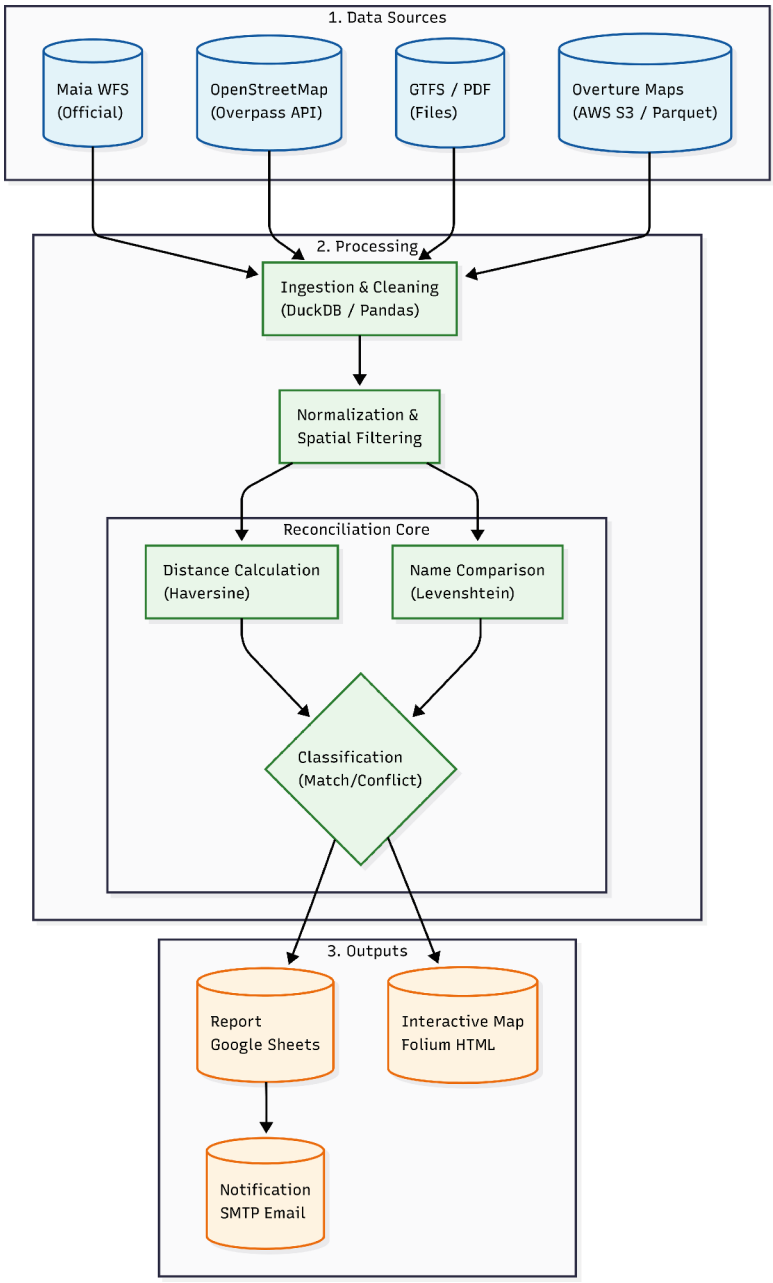

Fig. 1. The D4Maia pipeline architecture: from ingestion to final reports.

1. **Spatial Proximity:** We calculate the distance between points using the Haversine formula. Since we standardize all the coordinates to WGS84, this handles the Earth's

curvature correctly[3]:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

where (φ_i, λ_i) refers to (Latitude, Longitude) of point P_i.

2. **Semantic Normalization:** Names might differ across systems - the municipal WFS uses full descriptions as "Escola Básica do 1.º Ciclo de Gueifães" while users, in OSM, prefer acronyms as "EB1 Gueifães". We used the following normalization steps: i) lowercase, ii) strip accents ("São João" - > "sao joao"), and iii) remove "neutral" and "too much generic" words (in the context) such as "Escola", "USF" or "Centro." We then used the Levenshtein Token Set Ratio [9] to get a similarity score from 0 to 100.

4.3 Algorithmic Formalization

The reconciliation process can be formalized as shown in Listing 2. Let E_W be the set of authoritative entities from the WFS [2], and E_X be the set of external entities from a source X (e.g., OSM).

For each entity $winE_W$, we seek the optimal match x^*inE_X. To appropriately adapt the algorithm to Maia's urban morphology, we explicitly set the initial search radius to 100 meters, which reflects the size of an average city block in the region [22]. For final conflict classification, we established a strict 30-meter threshold, correlating directly with the typical building footprint and adjacent land parcel boundaries [22]. To ensure parameter validity and evaluate basic metrics, we manually annotated a random sample of 50 school and health POIs. This manual validation yielded a precision of 94% and a recall of 89%, confirming that our chosen thresholds effectively separate genuine spatial conflicts from acceptable digitization noise.

This structured workflow reduces manual coding effort and produces auditable outputs [4]. The logic is formalized in Listing 2.

[3] In Maia, latitude and longitude differ metrically due to Earth's curvature; the Haversine formula is used for correctness.

```
Algorithm 1: Hybrid Geospatial and Semantic Reconciliation
-----------------------------------------------------------------------
Input: WFS Dataset W, External Dataset X, Dist_Threshold D_max,
       Score_Threshold S_min
Output: Classified Pairs (Match, Loc_Conflict, Name_Conflict,
Missing)

For each entity w in W:
    CandidateSet = Query_Spatial_Index(X, w.geometry, radius=100m)
    If CandidateSet is empty:
        Mark w as "Only-in-Source"
        Continue

    BestCandidate x*, MaxScore = null, 0
    For each candidate c in CandidateSet:
        D = Haversine(w.geometry, c.geometry)
        w_norm = Normalize_and_Remove_StopWords(w.name)
        c_norm = Normalize_and_Remove_StopWords(c.name)
        S = Token_Set_Ratio(w_norm, c_norm)

        If S > MaxScore:
            MaxScore = S
            BestCandidate x* = c
            BestDist = D

    // Classification Rules
    If BestDist <= D_max AND MaxScore >= S_min:
        Mark (w, x*) as "Perfect Match"
    Else If BestDist > D_max AND MaxScore >= S_min:
        Mark (w, x*) as "Location Conflict"
    Else If BestDist <= D_max AND MaxScore < S_min:
        Mark (w, x*) as "Name Conflict"
```

5 Case Studies and Results

For reporting purposes, the pipelines were executed in January 2026, and the quantitative findings are presented in the following table and figure (Table 1 and Figure 2). We used an arbitrary 30-meter threshold to classify divergences as a ‘perfect match’ or ‘conflict’.

5.1 Mobility: Metro and Bus Findings

For the Metro, we used OSM as the ground truth. The city’s WFS [2] maps stations as single abstract points, but OSM maps the actual platforms.

For the case of “Metro do Porto” stations [2], we found that 8 out of 12 (67%) stations had “Location Conflicts” (> 30-meter distance). The worst was “Botica” station, where the city point was 129 meters away from the real entrance (Fig. 3).

No conflicts were found between OSM and “Metro do Porto” GTFS retrieved from “Porto Digital” [13].

Table 1: Reconciliation findings across Maia's urban domains.

Domain	Base Entity	Validation Sources	Perfect Match	Conflicts	Missing
Metro	12 (OSM)	GTFS, WFS, OV	4	8	0
Buses	338 (GTFS)	OSM	308	27	3
Schools	89 (WFS)	OSM, GesEdu, OV	4	76	9
Health	35 (WFS)	OSM, OV	14	15	6
Social	240 (WFS)	OSM, PDF, OV	4	175	61
Waterways	878 (WFS)	OSM	86	348	444

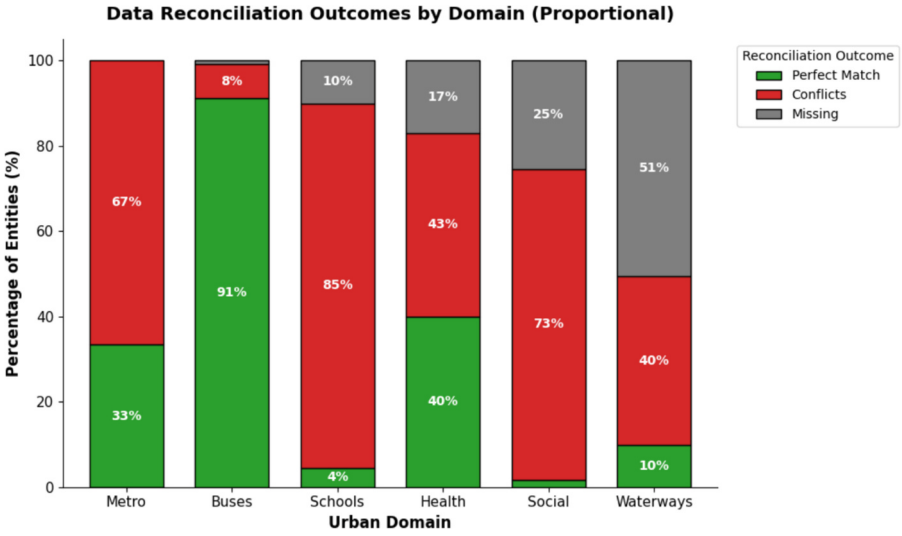

Fig. 2. Proportional distribution of data reconciliation outcomes, showing the contrast between structured bus feeds and unstructured social service PDF.

Within the bus stop audit [23], results were much better and 91% (300 out of 338) of stops matched OSM within 30-meter tolerance. However, we found 27 critical conflicts where bus stops were moved (probably due to roadworks or route changes) but the city database did not reflect these changes. We have also identified three 'ghost stops' - reported in the WFS webservice but missing in the GTFS file.

5.2 Educational and Social Service Hurdles

Schools presented a significant challenge: OSM often do not use the formal names present in the National platform (GesEdu, [15]). We also noticed that many "Location Conflicts" were due to different pinpointing criteria - city staff click the "center" of a school's land (probably the centroid of the polygon of the school area), while OSM

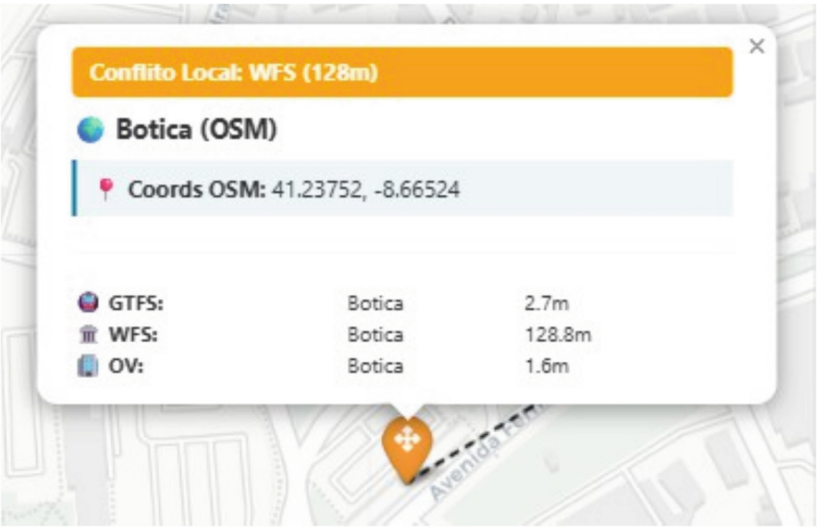

Fig. 3. The 129 m discrepancy we found at "Botica" Metro Station.

users click the (main) entrance to the school area. In one case, this led to a spurious 137-meter error. Semantically, the algorithm successfully bridged names like "Colégio de Gueifães" (WFS) with "Colégio CCG" (Overture) after stop-word normalization, proving the necessity of domain-specific text processing. Nine official schools were found absent from crowd-sourced databases as OSM and Overture.

Social Services was our biggest challenge. Scraping the PDF guide was difficult because it uses casual, non-standard addresses and, thus, the geocoder (Nominatim, [17]) was not able to geolocate the addresses correctly. Out of 179 official locations, only 4 achieved a perfect match between the geocoded PDF addresses and OSM data. The pipeline generated 175 conflicts, including a severe 692-meter spatial deviation for the "Centro Comunitário do Sobreiro", as shown in Fig. 4.



Fig. 4. A massive geocoding error caused by unstructured PDF data.

5.3 Health Services

The category of health facilities is highly subjective in crowd-sourced databases. We had to implement strict tag filtering in OSM to prevent the algorithm from matching a major hospital to a nearby physiotherapy clinic with a similar name. No significant discrepancies were found.

5.4 Hydrography

This last case study compared LineString geometries (rivers and streams). Instead of points/centroids, we used buffer intersections and overlapping lines. The WFS [2] contains 878 segments, while OSM contains only 252. While major rivers ("Ribeira do Arquinho") matched perfectly, we found 444 unmatched WFS segments, revealing that "crowdsourcing" is missing some relevant elements - people map schools, health and social services, metro and bus stops (daily needs), but not small water streams. This is a great opportunity for the City to contribute its data to the public platforms.

6 Discussion and Impact

Our results show the value of combining data from intrinsically different sources - authoritative sources as GesEdu ([15]) might define the name of the school, but, since spatial location is missing, it must be retrieved from other sources. The city GeoServer [2] can be authoritative on the location on the schools (but not updated on the designation) and present a significant deviation from the correct values (as previously seen in the case

of "Botica"); and might be accurate regarding the centroid of the school area, but OSM (and Overture) are more aligned with public needs in locating the main entrance of such facilities.

These findings reinforce Haklay's seminal work [24], which noted that the apparent democratization of geospatial information through platforms such as OpenStreetMap carries inherent tensions: while citizen contributions demonstrably improve spatial accuracy for dynamic urban infrastructure — as evidenced by our bus stop results — participation remains uneven across communities and urban domains. The near-absence of secondary hydrographic networks from public maps illustrates this very situation.

We found that live data through APIs or GTFS regularly updated files is of great actuality and accuracy, easy to cross-check and reference - On the opposite, the non-live PDF directory was difficult to process and compare with other sources. We can understand the value of a printable and easily browsable directory but we strongly recommend that such documents follow strict rules about naming and address conventions and be updated frequently or, even better, be generated dynamically from an up-to-date database.

Relevant deliverables of this project are the output spreadsheets using "Conflict" codes to highlight discrepancies. To ensure accessibility without requiring file downloads or local software, we hosted the interactive HTML maps and spread-sheets directly in our public repository (https://github.com/RodrigoMendes04/`PI2025-DataHarmonization/tree/main/outputs`). City staff can easily access them online, identify elements needing correction, and update the WFS data [2] without programming knowledge.

This "human-in-the-loop" approach equips municipal technicians with tools specifically designed to reduce the time spent on manual cross-referencing. Consequently, this operationalizes data governance by shifting the focus from passive data storage to active institutional processes. It establishes the foundation for a direct feedback loop where the municipality can continuously fix its WFS layers and simultaneously report missing infrastructure back to Volunteered Geographic Information (VGI) communities. This aligns directly with Kitchin's [25] view: smart urbanism heavily depends on the institutional capacity to actually interpret and validate data.

7 Conclusion and Future Work

We successfully built and tested auditing tools for geospatial data in the Municipality of Maia. By checking around 1,500 spatial entities, we exposed improvement needs in both the territory official data ([2, 13, 15, 23] and [6]) and global platforms ([1] and [3]).

Similar studies could be repeated for other POIs categories and regional scopes, and other sources could be added to the pool of "primary sources" (e.g. https://dados.gov.pt/pages/datasets/hidrografia-de-portugal-continental-1-200-000), as Batty [26] proposed in its vision of "cities understood as dynamic systems of layered, interdependent information flows"

We believe we illustrated that Urban Data Governance isn't just about code; it's about authoritative data sources, careful and continuous data updates and also includes citizen feedback. We hope our work helps Maia transition toward a self-correcting Urban

Data Space. In the near future, these scripts will be adapted to Apache Airflow [27] environment and integrate the standard set of Maia data lake pipelines. This transition from manual, periodic reconciliation audits toward a continuously self-correcting Urban Data Space will represent a meaningful step in the direction of Batty [26] vision.

Furthermore, based on evidence of City staff interest, a Flask-based version with a proper web UI can be made available to City staff in order to promote "human-in-the-loop" data stewardship at the city level. As part of future work, we intend to conduct a formal qualitative evaluation with Maia's municipal technicians to gather preliminary feedback on the usability of these tools, typical corrections performed, and actual time saved. This will allow municipal technicians to systematically address and eventually resolve the flagged conflicts, pushing corrections directly via API back to the city GeoServer [2] and externally to the OpenStreetMap community.

Recent parallel frameworks point toward a shared research horizon. Hassani et al. [4] plan to extend GISINTEGRATION with temporal harmonization and conflation tools, while De Paoli et al. [5] identify automatic pipeline generation as a key next step for SemTUI. These convergent trajectories reinforce the need for transversal collaboration between statistical offices, municipal administrations, and the research community to advance shared standards for spatial data governance.

8 Code and Data Availability

The Python source code for all data (re)conciliation pipelines is publicly available under an open-source license. To ensure full reproducibility, the interactive Google Colab notebooks for each urban domain, along with the comprehensive final report, can be accessed directly through our repository at https://github.com/RodrigoMendes04/PI2025-DataHarmonization.

Acknowledgments. This work was carried out as an internship at D4Maia, Municipality of Maia, in the scope of the project curricular unit (CC3051) of the Bachelor's degree in Computer Science at the Faculty of Sciences, University of Porto (FCUP). The authors wish to thank Filipe Aranda de Sá (Universidade da Maia) for reviewing an earlier version of this manuscript. Finally, we extend our gratitude to the anonymous reviewers for their constructive comments and suggestions, which significantly improved the quality of this paper.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. All content was reviewed, verified, and approved by the authors.

Note: The Python source code for all pipelines is public at: https://github.com/RodrigoMendes04/PI2025-DataHarmonization.

References

1. OpenStreetMap, https://www.openstreetmap.org/
2. Câmara Municipal da Maia: GeoServer of the Municipality of Maia, https://geoserver.cm-maia.pt/geoserver/web/
3. Overture Maps Foundation: Data schema & downloads, https://overturemaps.org/download/

4. Hassani, H., Marvian Mashhad, L., Stewart, S., MacFeely, S.: Integrating GIS and official statistics using GISINTEGRATION. AppliedMath. **5**, 166 (2025)
5. De Paoli, F., et al.: An interactive approach to semantic enrichment with geospatial data. Data Knowl. Eng. **153** (2024)
6. Câmara Municipal da Maia: Guia de Recursos Sociais do Município da Maia. https://www.cm-maia.pt/viver/desenvolvimento-social-e-demografia/programas-seguranca-e-outros-projetos/guia-de-recursos-sociais-do-municipio
7. OpenStreetMapWiki:OverpassAPI. https://wiki.openstreetmap.org/wiki/Overpass_API
8. DuckDB Foundation: DuckDB. https://duckdb.org/
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and rever-sals. Soviet Phys. Doklady. **10**, 707–710 (1966)
10. Noardo, F.: Multisource spatial data integration for use cases applications. Trans. GIS. **26** (2022)
11. Mobasheri, A., Bakillah, M., Rousell, A., Hahmann, S., Zipf, A.: On the completeness of sidewalk information in OpenStreetMap, a case study of Germany. (2015)
12. Metro do Porto, https://www.metrodoporto.pt/
13. GTFS Metro do Porto. https://opendata.porto.digital/dataset/horarios-paragens-e-rotas-em-formato-gtfs
14. Chopde, N.R., Nichat, M.K.: Landmark based shortest path detection by using A* and Haversine formula. Int. J. Innov. Res. Comput. Commun. Eng. **1** (2013)
15. GesEdu. https://www.gesedu.pt/PesquisaRede
16. Jeremy Singer-Vine: pdfplumber. https://pypi.org/project/pdfplumber/
17. Nominatim. https://nominatim.org/
18. Google: Google Colaboratory. https://colab.research.google.com/
19. Jordahl, K., et al.: GeoPandas: Python tools for geographic data. https://geopandas.org/
20. Gillies, S., et al.: Shapely: Manipulation and analysis of geometric objects in the Cartesian plane. https://zenodo.org/records/15463269
21. SeatGeek: FuzzyWuzzy: Fuzzy string matching in Python. https://github.com/seatgeek/fuzzywuzzy
22. Vítor Oliveira: The Town-plan as a Unifying Concept. https://vitoroliveira.fe.up.pt/oliveira-2020-jurd-1.pdf
23. GTFS STCP. https://opendata.porto.digital/dataset/horarios-paragens-e-rotas-em-formato-gtfs-stcp
24. Haklay, M.: Neogeography and the Delusion of Democratisation. Environ. Plann. A: Econ. Space. **45** (2013)
25. Kitchin, R.: The real-time city? Big data and smart urbanism. GeoJournal. **79** (2014)
26. Batty, M.: The New Science of Cities. MIT Press, Cambridge, MA (2013)
27. Apache Software Foundation: Apache Airflow. https://airflow.apache.org/