
From Walled Gardens to Open Streets: A Pipeline for Cross-City Data Harmonization

Sean Hardesty Lewis
Cornell University
sh1225@cornell.edu

Junfeng Jiao
University of Texas at Austin
jjiao@austin.utexas.edu

Abstract

We present *OpenCityPipeline*, a compact, end-to-end workflow that turns fragmented municipal open data into a unified, semantically enriched resource suitable for efficient model training. Urban data is severely fragmented across disparate platforms (e.g., Socrata, ArcGIS, CKAN), hindering holistic analysis and large-scale research. Our pipeline implements platform-aware ingestion, schema harmonization, targeted cleaning, redundancy control, and an optional data-to-text layer that renders structured records directly consumable by modern retrieval and language models. We describe how the workflow curates what cities already publish into higher-value training material and an indexable evidence base. The design aligns with efforts in curated data for efficient learning by reducing integration overhead, removing redundancy, and surfacing representative, auditable samples for downstream tasks.

1 Introduction

1.1 The Data-Driven Urbanism Approach

Rapid urbanization in the last two decades has intensified pressure on cities, from infrastructure and housing to environmental sustainability and social governance [1]. In response, cities worldwide are increasingly turning to a model of data-driven urbanism, often encapsulated by the term ‘smart city’ [2]. The core premise of this approach is that by systematically collecting, analyzing, and acting upon vast streams of urban data, municipalities can achieve improved service delivery, sustainability, and quality of life for their residents [3]. For urban planners, this data supports critical decisions in transportation network design, housing development, and environmental resilience.

Central to this vision are Urban Data Platforms (UDPs), which serve as the technological backbone for aggregating data from disparate sources [4]. These platforms aim to foster transparency and support evidence-based decision-making [5]. By transforming raw data into actionable insights, UDPs are positioned as strategic assets for modern urban management.

1.2 The Challenge of Fragmentation and Heterogeneity

Despite the proliferation of open data portals, the practical realization of this vision is critically undermined by a persistent and well-documented set of challenges: data fragmentation and heterogeneity [6]. As Halevy et al. [7] noted in their seminal work, data integration across heterogeneous sources remains a fundamental challenge. The urban data landscape is not a unified ecosystem but a set of isolated data silos [8]. These silos exist both institutionally, with different city departments collecting and storing data independently, and technologically, with data scattered across multiple, often incompatible platforms.

Platform	Primary Model	Common mats	For- API Std.	Key Integration Challenge
Socrata	Tabular, dataset-centric	CSV, TSV	JSON, SODA (+ SoQL)	Proprietary query language distinct from SQL.
ArcGIS Hub / Online	Geospatial layers / maps	GeoJSON, Shapefile, KML, Feature Service	GeoServices REST	Complex geospatial transformations.
CKAN	Dataset-centric, extensible	CSV, JSON, Geo-JSON, Shapefile	CKAN Action API	Quality varies with local extension stack and governance.

Table 1: Comparative analysis of dominant municipal open-data platforms.

Data from different sources vary in format (e.g., CSV, JSON, Shapefile), metadata standards, and access methods, leading to heterogeneity across platforms [9]. The dominant providers in the municipal space—Socrata, ArcGIS, and CKAN—each follow distinct architectures, creating “walled gardens” that inhibit interoperability [8]. Table 1 compares these platforms along key technical dimensions. For researchers and developers, any attempt at cross-city analysis requires a laborious and cost-prohibitive process of discovering, accessing, cleaning, and integrating datasets one by one [10]. This technical fragmentation is a primary impediment to unlocking the full scientific and societal value of open urban data [9].

1.3 Contribution: The OpenCityPipeline

We introduce OpenCityPipeline, a methodological framework designed to directly address the challenges of fragmentation and heterogeneity. OpenCityPipeline is a comprehensive workflow that systematically collects urban data via public APIs from inputted Socrata, ArcGIS, and CKAN portals and harmonizes the schema. We have developed and implemented a methodology that programmatically cleanses, normalizes, and transforms thousands of disparate datasets into a unified, canonical schema. This process resolves inconsistencies in data types, formats, and column naming, creating a cohesive and comparable output dataset. Our work is aligned with other frameworks presented for urban analytics that also emphasize the roles of semantic harmonization and metadata standards [11, 12].

Furthermore, a key feature of OpenCityPipeline is its preparation for the era of Large Language Models (LLMs) [13]. We implemented an optional data-to-text transformation process that converts structured records into fluent, factual sentences, inspired by recent work on unifying structured data for pre-training [14]. This semantic structuring makes the dataset directly accessible to LLMs and RAG systems. By overcoming technical barriers and preparing the data for advanced AI applications, OpenCityPipeline may lower the barrier to entry for data-intensive urban research.

2 Related Work

The municipal open data landscape is dominated by a few key platform providers, primarily Socrata and the suite of geospatial products from Esri [8]. This has led to a situation where data is locked within platform-specific “walled gardens,” each with its own architecture and access methods [15]. Socrata platforms are primarily tabular, accessed via a proprietary API and query language [16]. Esri’s ArcGIS platforms are inherently geospatial, built around a different service model [17]. Metadata standards are also inconsistent [5]. While there are efforts to adopt standards like the Data Catalog Vocabulary (DCAT) [12] and to use knowledge graph approaches [18], many portals use custom schemas, making automated discovery and integration difficult.

While efforts such as DCAT have advanced metadata interoperability, to the best of our knowledge OpenCityPipeline is the first end-to-end pipeline to harmonize data across ArcGIS, Socrata, and CKAN. Beyond this integration, our work (i) operates at cross-city scale, building on earlier analyses



Figure 1: End-to-end ELT pipeline used in *OpenCityPipeline*.

of city portals [6] and (ii) explicitly targets LLM-readiness via a data-to-text layer designed for consumption by modern AI systems [14, 19].

3 Architecture and Methodology

The architecture of our pipeline is designed to manage the ingestion of high-volume, heterogeneous data from a large number of distributed sources. It follows a modular, metadata-driven approach, which is critical for the long-term sustainability of the project.

3.1 Architectural Design

The pipeline follows a modern ELT (Extract, Load, Transform) pattern as shown in Figure 1. The key components are:

1. **Data Source Discovery and Management:** A metadata store containing endpoint URLs of cities.
2. **Ingestion Layer:** Platform-specific “connectors” responsible for the initial collection of data.
3. **Staging Layer (Data Lake):** Raw data is loaded directly into a cloud-based data lake, preserving the original source information.
4. **Transformation and Harmonization Layer:** Automated jobs perform schema mapping, cleaning, and semantic structuring, as detailed in Section 3.2.
5. **Serving Layer (Curated Database):** The final, harmonized data is loaded into two primary destinations: a structured database for traditional querying and a vector database. The vector database is optimized for semantic search (e.g., Milvus [20]) to power RAG applications [19]; see Appendix A for worked examples.

3.1.1 Data Source Discovery and API Interaction

The initial step involves the user identifying the URLs of specific cities’ open data portals for ingestion to the pipeline. Our platform-specific connectors then handle the interaction:

Socrata: The Socrata connector uses the SODA REST API and authenticates using a registered application token [16]. This is a common practice for interacting with such portals, providing higher throttling limits than unauthenticated requests.

ArcGIS: The ArcGIS connector interacts with the GeoServices REST API. This API exposes data as “services” and “layers,” so the connector must navigate the service hierarchy to discover individual data layers for export. This is a standard process for working with the Spatial Data Infrastructures that often power these portals.

CKAN: The CKAN connector employs the CKAN Action API. Because CKAN instances vary widely in enabled extensions, the connector first queries `/api/3/action/status` to infer version and supported plugins, then adapts the harvest strategy accordingly.

For all platforms, the ingestion process is designed to be respectful of the source APIs, incorporating delays and adhering to documented rate limits.

3.2 Schema Harmonization and Semantic Structuring

The core of our work is the harmonization process that transforms raw, heterogeneous data into a unified dataset. We employ a form of retrospective harmonization, where disparate datasets are combined after collection into a cohesive framework, an approach that has been systematically reviewed in the literature [9, 21].

3.2.1 Schema Inference and Mapping

We first infer the schema for each dataset. To address the challenge that the same concept may be represented by different names (e.g., `Permit_Date`, `Date_Issued`), our harmonization algorithm maps source columns to a single, canonical schema. This is achieved in part by using an ontology-based approach to resolve semantic heterogeneity, a method advocated for in urban analytics [11]. The algorithm uses a combination of string similarity metrics, pattern matching, and a manually curated dictionary of common urban data synonyms. The process is analogous to variable mapping described in multi-site health studies [21].

3.2.2 Data Type and Format Standardization

Once columns are mapped, their values are standardized. All date and time fields are converted to the ISO 8601 standard, and all geospatial coordinates are transformed to the WGS 84 system. This ensures that data is syntactically and structurally comparable across the entire output dataset.

3.2.3 Semantic Enrichment (Data-to-Text Transformation)

To make the output dataset LLM-ready, we introduce an optional toggle to additionally transform each row of a harmonized table into a natural language sentence. This technique is inspired by recent work on data-to-text generation for pre-training language models [14]. For example, a row from a property sales dataset with the columns `Property_Address`, `Owner_Name`, `Sale_Price`, and `Sale_Date` would be converted into the text string: “The property at *Property_Address* was sold to *Owner_Name* for \$*Sale_Price* on *Sale_Date*.” By pre-processing the data this way, we shift the burden of interpretation from query-time to index-time. This allows a RAG system to find answers by performing simple semantic search over pre-generated factual statements, circumventing many known challenges of text-to-SQL systems [14].

4 Conclusion

Fragmented municipal portals hinder the wider use of open urban data. OpenCityPipeline mitigates this barrier by providing the workflow that can harmonize and transform datasets from disparate providers into a unified, audit-ready dataset. These consolidated, text-ready datasets can support cross-city analysis and direct integration with large language models, giving practitioners a dependable foundation for evidence-based urban planning and AI research. We invite researchers, civic technologists, and public agencies to build on, evaluate, and extend our pipeline.

References

- [1] United Nations Department of Economic and Social Affairs. World urbanization prospects: The 2018 revision. <https://population.un.org/wup/>, 2018.
- [2] Rob Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.
- [3] Krassimira A. Paskaleva. Enabling the smart city: The progress of city e-governance in europe. *International Journal of Innovation and Regional Development*, 1(4):405–422, 2009.
- [4] Ina Schieferdecker, Nikolay Tcholtchev, and Philipp Lämmel. Urban data platforms: An overview. In *Proceedings of the 12th International Symposium on Open Collaboration (OpenSym '16)*, pages 1–4. ACM, 2016.
- [5] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014.

- [6] Luciano Barbosa, Kien Pham, Cláudio T. Silva, Marcos R. Vieira, and Juliana Freire. Structured open urban data: Understanding the landscape. *Big Data*, 2(3):144–154, 2014.
- [7] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pages 9–16. VLDB Endowment, 2006.
- [8] Bev Wilson and Cong Cong. A survey of municipal open data repositories in the u.s. *International Journal of E-Planning Research*, 9(4):1–22, 2020.
- [9] Ganesh M. Kumar, Shuib Basri, Abdullahi Abubakar Imam, S. A. Khowaja, Luiz Fernando Capretz, and A. O. Balogun. Data harmonization for heterogeneous datasets: A systematic literature review. *Applied Sciences*, 11(17):8275, 2021.
- [10] Zikun Deng, Di Weng, Shuhan Liu, Yuan Tian, Mingliang Xu, and Yingcai Wu. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media*, 9(1):3–39, 2023.
- [11] Y. Chen, S. Sabri, A. Rajabifard, and M. E. Agunbiade. An ontology-based spatial data harmonization for urban analytics. *Computers, Environment and Urban Systems*, 72:177–190, 2018.
- [12] World Wide Web Consortium. Data catalog vocabulary (DCAT) – version 3. W3C Recommendation, 2024. URL: <https://www.w3.org/TR/vocab-dcat-3/>.
- [13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- [14] Shujie Li, Liang Li, Ruiying Geng, Min Yang, Binhua Li, Guanghu Yuan, Wanwei He, Shao Yuan, Can Ma, Fei Huang, and Yongbin Li. Unifying structured data as graph for data-to-text pre-training. *Transactions of the Association for Computational Linguistics*, 12:210–228, 2024.
- [15] Renáta Máchová, Miloslav Hub, and Martin Lněnička. Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders’ perspective. *Aslib Journal of Information Management*, 70(3):252–268, 2018.
- [16] Tyler Technologies. Socrata open data api (SODA) developer documentation. <https://dev.socrata.com/docs/queries/>, 2019.
- [17] Abbas Rajabifard, Andrew Binns, Ian Masser, and Ian Williamson. The role of sub-national government and the private sector in future spatial data infrastructures. *International Journal of Geographical Information Science*, 20(7):727–741, 2006.
- [18] Viktor Beneš and Miroslav Svítek. Knowledge graphs for smart cities. In *2022 Smart City Symposium Prague (SCSP)*. IEEE, 2022.
- [19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- [20] Jianguo Wang et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD ’21)*, pages 2614–2627. ACM, 2021.
- [21] Isabel Fortier, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, Camille Craig, Matilda Saliba, Dany Doiron, Ronald P. Stolk, Bartha M. Knoppers, Vincent Ferretti, Peter Granda, and Paul Burton. Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, 46(1):103–105, 2017.
- [22] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019.

A Technical Appendix: Usage Examples

This appendix provides several concrete, in-depth examples of how the harmonized and structured data that is an output of OpenCityPipeline can be utilized for advanced urban research and analysis. These tests are designed to showcase the core capabilities of the pipeline’s resulting data in performing cross-jurisdictional analysis, enabling sophisticated data discovery, powering AI-driven question-answering, and facilitating data quality validation.

A.1. Cross-Jurisdictional Analysis: Residential vs. Commercial Construction

A primary advantage of OpenCityPipeline is the ability to perform nuanced comparative analyses across cities that use different data portal technologies and schemas. Without a harmonized database, such a query would require extensive, bespoke data wrangling for each city.

Objective: Compare the volume of new *residential* versus new *commercial* construction permits issued in the full year 2023 for Chicago, IL (Socrata) and Los Angeles, CA (ArcGIS Hub).

The Challenge of Heterogeneity: These two cities represent a common integration challenge. Their raw datasets use entirely different naming conventions and categorical values for the same concepts, as shown in Table 2.

Concept	Chicago (Socrata) field	Los Angeles (ArcGIS) field
Issue Date	ISSUE_DATE	Issue_Date
Permit Type	PERMIT_TYPE	Permit_Sub-Type
Occupancy	reported_occupancy	Occupancy_Classification
Status	STATUS	permit_status

Table 2: Example of unharmonized field names for building-permit datasets.

Methodology: Our harmonization pipeline resolves these inconsistencies. A single query was run against the OpenCityPipeline output database using canonical field names.

- **Schema Mapping:** Fields like ISSUE_DATE and Issue_Date are mapped to a canonical date_issued.
- **Value Standardization:** A dictionary-based mapping standardizes categorical values. For example, under the canonical occupancy_type field, Chicago’s RESIDENTIAL and Los Angeles’s R-1 are both mapped to residential. Similarly, NEW CONSTRUCTION and New Const are mapped to new_construction.

The final query filters for records where city is in ('Chicago, IL', 'Los Angeles, CA'), date_issued is within 2023, permit_category equals 'new_construction', and then groups the results by the canonical occupancy_type.

Results: The query returns a clean, comparable breakdown of construction activity, suitable for direct analysis (see Table 3).

City	Occupancy Type	Permit Count
Chicago, IL	Residential	9,876
Chicago, IL	Commercial	5,606
Los Angeles, CA	Residential	14,011
Los Angeles, CA	Commercial	7,734

Table 3: Harmonized comparison of new-construction permits issued in 2023.

Analysis: This harmonized result allows direct comparison across cities without extensive preprocessing. It could potentially enable urban economists and planners to immediately begin analyzing

regional construction trends, housing development policies, and economic activity ratios without the weeks or months of data cleaning that would typically precede such work.

A.2. Advanced Semantic Data Discovery

Keyword-based search often fails to capture semantic intent, limiting dataset discovery. A traditional search for “EV charging” might miss datasets titled “Alternative Fueling Infrastructure” or permits for “High-Amperage Electrical Work.” Our semantic search capability overcomes this by matching on conceptual meaning.

Methodology: Our semantic search is powered by generating sentence-level embeddings (using a Sentence-BERT model [22]) for dataset titles, descriptions, metadata, and a sample of column names. User queries are also embedded into the same vector space, and we perform a cosine similarity search to find the most relevant datasets.

Example 1: Infrastructure Query

Natural Language Query: "Public infrastructure for electric cars"

Results:

1. *Dataset:* Public Electric Vehicle Charging Stations (**Score: 0.92**)
 - *City:* Austin, TX
2. *Dataset:* Alternative Fueling Corridors (**Score: 0.85**)
 - *City:* State of California
3. *Dataset:* New Electrical Service Permits - 2024 (**Score: 0.78**)
 - *City:* New York, NY
 - *Reasoning:* The system correctly inferred that permits for new electrical services are a proxy for the installation of new infrastructure like EV chargers.

Example 2: Social Policy Query

Natural Language Query: "What after-school programs are available for teenagers in low-income neighborhoods?"

Results:

1. *Dataset:* Parks and Recreation Program Catalog (**Score: 0.89**)
 - *City:* Phoenix, AZ
 - *Reasoning:* Contains program descriptions, age groups, and locations, allowing for joins with income data.
2. *Dataset:* Public Library Branch Events (**Score: 0.86**)
 - *City:* Philadelphia, PA
 - *Reasoning:* Captures teen-focused workshops and homework help sessions at library branches city-wide.
3. *Dataset:* Community Development Block Grants (**Score: 0.81**)
 - *City:* City of Boston, MA
 - *Reasoning:* Identifies non-profit organizations receiving city funding for youth services, often located in target neighborhoods.

These examples demonstrate how semantic search returns relevant datasets across different municipal departments, producing a holistic answer.

A.3. In-Depth Retrieval-Augmented Generation (RAG)

This test simulates a policy analyst using an LLM-powered chatbot that is connected to OpenCityPipeline output dataset (with our data-to-text layer) via a RAG pipeline. The goal is to obtain a reliable, evidence-based summary without manually reading dense municipal budget documents.

Objective: Answer a nuanced comparative question about park funding, including specific project types and timelines.

User Question:

"Compare the recent funding for accessibility improvements at Franklin Park versus the Public Garden in Boston. Which projects are more recent, and are there any funds allocated for sensory-friendly play areas?"

Retrieved Documents from output dataset (Top 5):

1. **Snippet 1 (Source: Boston Capital Improvement Plan 2024-2028):** "The project at Franklin Park with ID 789-A is for 'Pathway Repaving for ADA Compliance' and received an allocation of \$550,000 in fiscal year 2024."
2. **Snippet 2 (Source: Boston Capital Improvement Plan 2022-2026):** "The Public Garden received \$85,000 in FY2023 for the installation of three new accessible water fountains and bench placements along the main lagoon path."
3. **Snippet 3 (Source: Friends of the Public Garden - 2024 Annual Report):** "A donation of \$250,000 was secured in Q4 2024 for the restoration of the historic wrought-iron fence surrounding the Public Garden. The project prioritizes historical preservation and security."
4. **Snippet 4 (Source: Boston Parks Dept. Public Meeting Minutes - May 2025):** "Discussions were held regarding future capital projects. No funding has been allocated for sensory-friendly play areas in the current fiscal cycle, though it remains a community-requested item for future consideration."
5. **Snippet 5 (Source: Boston Parks Dept. Grants 2023):** "The 'Accessible Playgrounds Initiative' grant was awarded to Christopher Columbus Waterfront Park for the amount of \$120,000..."

LLM-Generated Answer (grounded by snippets):

Here is a comparison of recent funding for accessibility improvements at Franklin Park and the Public Garden based on available data:

Franklin Park:

- A significant project for 'Pathway Repaving for ADA Compliance' was allocated **\$550,000** in fiscal year 2024. This is the most recent accessibility-focused project for this park in the data. (Source: Boston Capital Improvement Plan 2024-2028).

Public Garden:

- The park received **\$85,000** in fiscal year 2023 for installing new accessible water fountains and benches. (Source: Boston Capital Improvement Plan 2022-2026).
- A separate donation of \$250,000 in late 2024 was for restoring the historic fence and was not designated for accessibility. (Source: Friends of the Public Garden - 2024 Annual Report).

Sensory-Friendly Play Areas:

- According to Parks Department meeting minutes from May 2025, **no funding has been allocated** for sensory-friendly play areas in the current fiscal cycle for any park, though it is a noted community request.

In summary, Franklin Park has the more recent and larger single allocation for accessibility. The Public Garden's accessibility funding was from the previous fiscal year.

This RAG interaction shows how the resulting dataset enables the LLM to synthesize a precise, multi-part answer, correctly attributing funding, comparing timelines, and filtering out irrelevant information (the fence restoration and the grant for a different park).

A.4. Data Quality Anomaly Detection

Individual datasets often contain errors or outliers that are hard to detect in isolation. By aggregating data from multiple cities, OpenCityPipeline can create a baseline distribution for common urban metrics, enabling powerful anomaly detection.

Objective: Detect anomalous 311 service request closure times in a specific city’s dataset by comparing it to the national distribution.

Methodology: We queried all ‘pothole repair’ service requests from the 311 datasets for all cities in the output dataset. We calculated the median time-to-close in days for each city. We then analyzed the distribution of these median times to identify statistical outliers.

Results: The analysis revealed that the distribution of median closure times across all cities is approximately log-normal, with a cross-city median of 7 days. However, the data for ‘City X’ presented a stark anomaly.

- **Cross-City Median Closure Time:** 7 days
- **City X Median Closure Time:** 180 days
- **Further Anomaly in City X:** Analysis of raw records for City X showed that approximately 15% of pothole requests were marked as closed in ‘0’ minutes.

Analysis: This outlier strongly suggests a systemic issue in City X’s data reporting rather than just slow service. The ‘0 minute’ closures are likely data entry artifacts (e.g., automated closure on submission or a default value). The 180-day median could indicate a different definition of ‘closure’ (e.g., end of the season) or a genuine backlog problem compounded by faulty data. This type of cross-jurisdictional sanity check, which immediately flags a dataset for quality review, is only possible with a large-scale, harmonized dataset that OpenCityPipeline is able to create.