Developing an Urban Gazetteer

A Semantic Web Database for Humanities Data

Vincent Ducatteeuw
Department of History
Ghent University
Ghent, Belgium
vincent.ducatteeuw@ugent.be

ABSTRACT

This talk discusses the development of a spatiotemporal data model for an urban gazetteer. The function of gazetteers is to obtain descriptions uniquely identifying places referred to in discourse. Often, they are lists of places containing place name, feature type and geographical extent. Contemporary digital gazetteers (e.g. World Historical Gazetteer and Pleiades) are valuable tools for geographical knowledge of the past and the structuring of humanities data. However, scholars and GLAM (Galleries, Libraries, Archives and Museums) specialists often require information about entities on an intra-city scale. This presentation explores the model and implementation of an urban gazetteer using CIDOC CRM as a top-level ontology. The model will closely follow international gazetteer standards (i.e. Linked Places Format) in order to ensure interoperability with other gazetteer datasets. To move towards a FAIR (Findable, Accessible, Interoperable, and Reusable) approach, humanities data from the urban gazetteer will be published as Linked Open Data (LOD) and searchable via (Geo)SPARQL.

CCS CONCEPTS

• Information systems → Data management systems → Database design and models → Graph-based database models; • **Information systems → Information systems applications → Spatial-temporal systems → Geographic information systems**; • Information systems → Information systems applications → Digital libraries and archives; • **Information systems → Information retrieval → Document representation → Ontologies**

KEYWORDS

digital humanities, spatial humanities, spatial history, urban history, spatiotemporal analysis, gazetteer development, urban gazetteer, semantic technologies, linked open data, modelling geohistorical data, CIDOC CRM, GeoSPARQL

1 What is an Urban Gazetteer?

Gazetteers often serve as important research tools in (digital) humanities. They act as specialized, authoritative indexes for places and are used to retrieve geospatial information. Over the last decades, the link between the semantic and spatial domain has been formalized into digital gazetteers [1]. This formalization was both enabled and necessary for the transfer of knowledge from traditional publications to the web. It resulted in gazetteers functioning as digital knowledge organization systems (KOS) arranging information by geographic coordinates and semantic relations. However, many existing digital gazetteers are not granular enough for researchers and GLAM specialists interested in urban history. Historical sources, such as census data, address books and tax registries, would be more accessible when linked to digital gazetteers with a higher level of granularity. This would, for example, involve the creation of URIs (Uniform Resource Identifier) for streets as historical places throughout time, which help to map historical city directories data [2]. It would seem that urban gazetteers are the answer to the granularity problem, but it is striking that there are only a few examples of them, despite the growing interest (e.g. MoEML [3]).

1.1 Design Challenges

What explains the low number of urban gazetteers when compared to other types of gazetteers? On the surface, urban gazetteers and large scale gazetteers look quite similar. They both deal with places and their geotemporal properties. Consequently, they face similar modelling problems related to place disambiguation, hierarchy and place typology. But when

studied further it seems that higher levels of granularity come with more complex variants of these issues. Take for example place disambiguation, a concept that refers to the issue of place identity. In order to assign URIs to places, a concept of place should be defined together with explicit identity criteria. These criteria determine what a place is, when it begins and ends. This is relevant because digital gazetteers require a formalization of place in order to extract and model place data. As places become more granular, the need to disambiguate based on these criteria only increases [4].

Unfortunately, theories about identity criteria that take into account database management are sparse. Those that do exist might not be fully applicable to urban places as these *behave* in a different way. This behavior is determined by the ontological assumptions researchers make when deciding if a place maintains or loses its identity through change [5]. A definition of place is an important part of these assumptions. However, place is often not explicitly defined in GIS research nor is there a consensus on how to deal with it [6]. Researchers interested in place do often refer to human geographical interpretations of the concept, defining place as space made meaningful. A full discussion about space and place would go beyond the scope of this talk, but it is important to keep these comments in mind as place concepts influence the structure and content of (urban) gazetteers.

Despite the differences, it is still worthwhile to view urban gazetteers as a subcategory of large scale (settlement) gazetteers due to the fact that some methodological and technical challenges overlap. Large scale gazetteers should therefore be studied to learn how they deal with the challenges of place. In this presentation the design of the Linked Places Format (LPF) is discussed as it is a common interconnection standard for gazetteer datasets [7, 8]. LPF provides a uniform way for linking place data coming from multiple projects. The format acts as a template for both the WHG platform and Pelagios' Recogito and Peripleo-lite software. However, it is important to note that LPF does not aim to be the only way of representing historical place data. Depending on the (historical) sources and requirements of the project the resulting data will be very distinct. As this is also the case for the urban gazetteer, LPF cannot be fully adopted. Close alignment with LPF is still very valuable to ease interoperability with other gazetteers, including WHG.

LPF is designed around the JSON-LD format and provides (optional) temporal scoping of place data through the use of GeoJSON-T [9]. GeoJSON-T is an experimental extension that standardizes the representation of temporal properties in GeoJSON. The extension provides a solution to the issue of modelling spatiotemporal data, as space and time are often managed as separate entities in GIS [10]. When working with gazetteers, space and time must form a coherent continuum, aptly called spacetime. A Linked Places record takes the form of a GeoJSON-LD FeatureCollection containing one or more Feature objects (i.e. places). For each place, LPF accommodates multiple names, types, geometries, relations, descriptions, depictions and links. This is an important feature as this enables

a spatiotemporal framework, necessary for historical gazetteers. Only through this framework is it possible to uniquely describe a place by locating it in spacetime, even as the place changes over time [11]. In LPF, a 'when' element from GeoJSON-T can be used to temporally scope an entire place, name, geometry, type or relation. If a 'when' element is used, it must include one or more timespan which allow to accurately date place information. As of present, terms in LPF are either defined using existing ontologies or they need to be defined in a Linked Pasts Ontology (lpo:), which is an ongoing process [12].

2 Proposed Model for an Urban Gazetteer

LPF facilitates several of the challenges digital gazetteers face in modelling (geospatial) gazetteer data. The model proposed here attempts to closely align with the structure of LPF in order to ease interoperability but it is defined using classes from the CIDOC Conceptual Reference Model (CRM) - unlike LPF [13]. CRM is an upper level ontology for the semantic integration of cultural heritage data. It was conceived in 1996 to support the management and interoperability of museum collections metadata, but the ontology is broad enough to model other topics. CRM is currently maintained by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) and has been an accepted ISO standard since 2006 [14]. An extension for geoinformation called CRMgeo introduced the concept of 'Spacetime Volume', which is essential for modelling spatiotemporal properties [15]. Since CRM version 6.2.2 some features of CRMgeo have become part of the CRM core. The model of a CRM-based gazetteer for medieval and early modern places has been discussed by Schneider et al. [16]. The urban gazetteer presented here modifies their reasoning centered around the CRM concept of spacetime volume to further align it with LPF and CRM version 7.1.1 (Figure 1) [17].
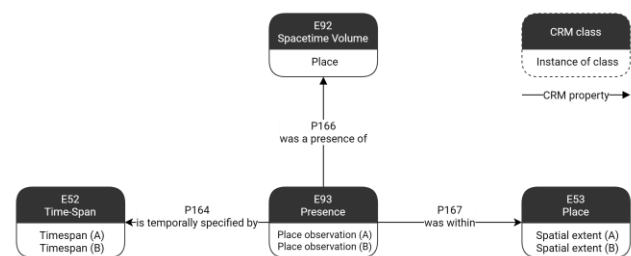
Figure 1: simplified view of the model showcasing the CRM classes and properties used to link multiple historical observations (E93 Presence) to a place (E92 Spacetime Volume).

The model is centered around the CRM class 'E92 Spacetime Volume' and its subclass 'E93 Presence'. CRM 7.1.1 documentation states:

"*E92 Spacetime Volume is used to document geometric extents in the physical spacetime containing actual or possible positions of*

things or happenings, in particular in those cases when the changes of place to be documented cannot be reduced to distinct events, because the spatial extent changes continuously." [17]

Using E92 together with associated properties and classes, it is possible to model the extent of (physical) places in space at any moment of their existence (i.e. model a gazetteer). Spatiotemporal properties of places can be modelled using properties 'P160 has temporal projection: E52 Time-Span' and 'P161 has spatial projection: E53 Place'. The property P160 associates a spacetime volume with the temporal extent it covers over the geographical area of its definition, while P161 connects a spacetime volume with the spatial extent of the phenomenon during its predetermined timespan.

In the model presented here, place data are derived from source observations, similar to attestations in LPF. Observations act as the basic unit to convey spatiotemporal data for the urban gazetteer. They are modelled as instances of CRM class 'E93 Presence'. E93 is a subclass of E92 and can be understood as a 'snapshot' of a spacetime volume (i.e. the spatial extent of a place at a fixed point in time). Multiple observations of the same place can be grouped using the property 'P166 was a presence of: E92 Spacetime Volume' as shown in figure 1. The spatial extent of an instance of E93 can be defined by the property 'P167 was within: E53 Place'. This property associates an instance of 'E53 Place' with the spatial projection of a place observation. The temporal extent of a place observation can be specified by 'P164 is temporally specified by: E52 Time-Span'. In the model presented here, both 'E52 Time-Span' and E53 Place are interpreted as declarative, instead of phenomenal. For example, the timespan of a place observation could be mentioned in the source or implicitly derived through dating of the source. In both cases, this is a human declaration approximating phenomena in the world. CRMgeo introduced this differentiation to clarify that unique properties of spacetime in the real world (phenomenal) are not the same as information describing that world (declarative) [15].
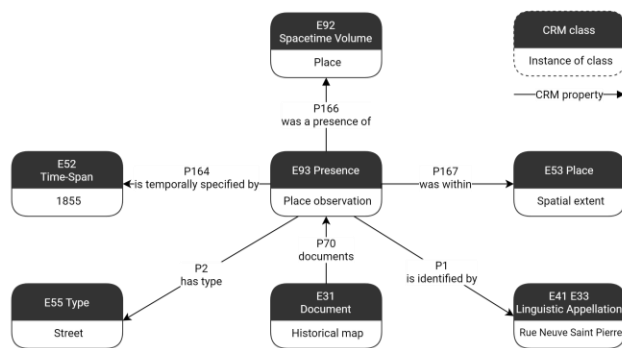

Figure 2: simplified view of the model illustrating how place information belongs to a unique presence (E93) and is derived from one source (E31).

Place information can be modelled using properties linked to E93. Figure 2 shows spatial extent and a French toponym

obtained from a historical map for the Sint-Pietersnieuwstraat in Ghent. Every piece of information that is derived from the source is part of a unique presence. This is a crucial principle for the proposed gazetteer. If another historical source mentions additional place information, a new presence is created.

As mentioned, the spatial extent of a place can be defined using E53 Place. Topological relations between places at fixed moments in time can be modelled using properties associated with E53: 'P89 falls within: E53 Place', 'P121 overlaps with: E53 Place', 'P122 borders with: E53 Place'. Property 'P168 place is defined by: E94 Space Primitive' allows to link CRM encoded data to geometries used in GIS. An instance of 'E94 Space Primitive' defines an instance of E53 Place where the identity of the place is derived from its geometric definition (e.g. a geometry serialized in WKT format) To maintain compatibility with the Open Geospatial Consortium (OGC) GeoSPARQL standards, certain CRM classes are mapped to GeoSPARQL classes (Figure 3) [15]. This allows the model to represent and query geospatial RDF data through the use of GeoSPARQL. For example, inferring topological relations between places based on geometry data.
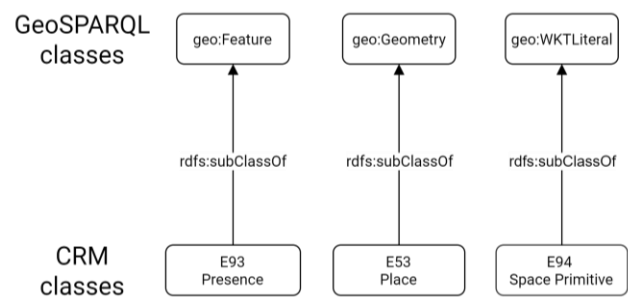
Figure 3: mapping of CRM classes to GeoSPARQL classes.

A place typology can be modelled in CRM using 'P2 has type: E55 Type' as E55 Type is CRM's way to represent domain specific ontologies and controlled vocabularies. WHG has created a subset list of 165 terms from the Getty Art and Architecture Thesaurus (AAT) specifically for gazetteers [18]. The hierarchy implemented in this AAT subset can be modelled using 'P127 has broader term: E55 Type of E55 Type'. Place names can be linked to an E93 Presence via the property 'P1 is identified by: E41_E33_Linguistic_Appellation'. The particular language of a place name can be declared with the property 'P72 has language: E56 Language'.

3 Implementation and Future Challenges

The data model is currently being tested using a corpus of datasets about streets in Ghent. These were mapped to the model and converted to RDF Turtle format via an RDF extension to OpenRefine [19]. Subsequently, the graph data was uploaded on TriplyDB, an RDF Triplestore supporting GeoSPARQL [20]. Data can be queried through the Yasgui SPARQL editor which provides advanced features for queries and results. After

additional development, the urban gazetteer will be used in two historical projects.

Gent Gemapt (2020-2023) is a public history project building a presentation platform for cultural heritage collections focused on the city of Ghent [21]. The main premise is that various heritage collections can be made accessible - for both researchers and the general public - through historical maps. The end goal is to provide a multidimensional map of Ghent that can give a better understanding of the urban landscape. An ongoing process is to explore how the urban gazetteer can be used to organize the different urban places.

Collective Action Belgium, is a case study led by the Ghent Centre for Digital Humanities. It focuses on social history in the 20th century and aims to trace the dynamics of collective action in Belgium as reported in Belgian newspapers. The research scenario is part of DATA-KBR-BE (2020-2022), an interdisciplinary research collaboration to optimize data-level access to KBR's (The Royal Library of Belgium) digitized collections for humanities research [22]. The aim is to integrate extracted data (i.e. places, actors and events) in the urban gazetteer. The mapping of actors and events will require an extension to the present model. This seems feasible as other humanities projects are already using the CRM ontology to map actors and events [23, 24].

Yet, many challenges remain to improve the urban gazetteer as a tool for humanities research. The next step is to define identity criteria for urban places and create a place disambiguation method. This is crucial for the aggregation of place observations and the creation of place URIs. Feature type will most likely be one of the main criteria to identify a place, which means a type vocabulary for urban places needs to be constructed. The urban typology should align with existing vocabulary standards (e.g. AAT) where possible, implementing additional types only when necessary. For urban place concepts it might be interesting to look at the usability of heritage types developed by the Flanders Heritage Agency [25]. Furthermore, it is also possible that researchers have different interpretations of the same source. At the moment the gazetteer cannot adequately implement different assertions of the same source.

To reach a critical mass of place data, a workflow needs to be designed for the integration of new sources. The use of the International image Interoperability Framework (IIIF) seems promising for this, as the IIIF Maps Technical Specification Group is working on an extension for georeferencing and geotagging IIIF images [26]. Allmaps, a project under development by Bert Spaan, can georeference IIIF maps and can be combined with Annotorious to annotate places visible on those maps [27]. Annotations on IIIF resources could be stored in the gazetteer database making it possible to query place data and retrieve a visual link to the original source material.

REFERENCES

[1] Humphrey Southall, Ruth Mostern and Merrick Berman. 2011. On Historical Gazetteers. *International Journal Of Humanities And Arts Computing* 5, 2 (2011), 127-145. https://doi.org/10.3366/ijhac.2011.0028

[2] Isabella di Lenardo, Raphaël Barman, Albane Descombes and Frédéric Kaplan. 2019. Repopulating Paris: massive extraction of 4 Million addresses from city directories between 1839 and 1922. *DH2019* (2019). https://doi.org/10.34894/MNF5VQ

[3] University of Victoria. 2021. Map of Early Modern London. Retrieved August 8, 2021 from https://mapoflondon.uvic.ca/

[4] Janelle Jenstad. 2020. Janelle Jenstad on Urban Gazetteer Design. Video. (9 September 2020). Retrieved August 8, 2021 from https://medium.com/pelagios/janelle-jenstad-on-urban-gazetteer-design-698852e55bea

[5] Pawel Garbacz, Bogumił Szady and Agnieszka Ławrynowicz. 2021. Identity Of Historical Localities. *Applied Ontology* 16, 1 (Jan, 2021), 55-86. http://dx.doi.org/10.3233/AO-200235

[6] Daniel Wagner, Alexander Zipf and Rene Westerholt. 2020. Place in the GIScience Community – an Indicative and Preliminary Systematic Literature Review. *Second International Symposium on Platial Information Science (PLATIAL'19).* Coventry, United Kingdom. https://doi.org/10.5281/zenodo.3628855

[7] Karl E. Grossner, Merrick Berman, Rainer Simon. 2017. Linked Places: A Modeling Pattern and Software for Representing Historical Movement. *DH2017* (2017).

[8] Karl Grossner. 2021. The Linked Places Format (LPF). Retrieved August 18, 2021 from https://github.com/LinkedPasts/linked-places-format

[9] Karl Grossner. 2019. GeoJSON-T. Retrieved August 18, 2021 from https://github.com/kgeographer/geojson-t

[10] Christophe Claramunt Siabato, Sergio Ilarri and Miguel Angel Manso-Callejo. 2018. A Survey of Modelling Trends in Temporal GIS. *ACM Computing Surveys*, 51, 2, Article 30 (June 2018), 41 pages. https://doi.org/10.1145/3141772

[11] Ryan Shaw. 2016. Gazetteers Enriched: A Conceptual Basis for Linking Gazetteers with Other Kinds of Information. In *Placing Names: Enriching and Integrating Gazetteers* (eds. Berman et al.), Indiana University Press, Bloomington. https://doi.org/10.2307/j.ctt2005zq7

[12] Karl Grossner. 2020. Linked Pasts Ontology (:lpo). Retrieved August 22, 2021 from https://github.com/LinkedPasts/linked-pasts-ontology

[13] CIDOC: International Committee for Documentation. 2021. CIDOC CRM: Conceptual Reference Model. Retrieved August 22, 2021 from http://www.cidoc-crm.org/

[14] ISO. 2020. *Information and documentation - A reference ontology for the interchange of cultural heritage information* (Standard No. 21127:2014). Retrieved August 22, 2021 from https://www.iso.org/standard/57832.html

[15] Gerald Hiebel, Martin Doerr and Øyvind Eide. 2016. Crmgeo: A Spatiotemporal Extension Of CIDOC-CRM. *International Journal On Digital Libraries* 18, 4 (2017): 271-279. https://doi.org/10.1007/s00799-016-0192-4

[16] Philipp Schneider, Jim Jones, Torsten Hiltmann and Tomi Kauppinen. 2021. Challenge-Derived Design Practices For A Semantic Gazetteer For Medieval And Early Modern Places. *Semantic Web* 12, 3 (2021): 493-515. https://doi.org/10.3233/SW-200394

[17] CIDOC CRM Special Interest Group. 2021. Volume A: Definition of the CIDOC Conceptual Reference Model (Version 7.1.1). Retrieved August 26, 2021 from http://www.cidoc-crm.org/version/version-7.1.1

[18] Linked Places Format. 2021. Feature types AAT. Retrieved 20 August, 2021 from https://github.com/LinkedPasts/linked-places-format/blob/master/feature-types-AAT_20210118.tsv

[19] Stkenny. 2021. Grefine-rdf-extension. Retrieved 05 August, 2021 from https://github.com/stkenny/grefine-rdf-extension

[20] TriplyDB. 2021. Retrieved 10 August, 2021 from https://triplydb.com/

[21] Ghent Centre for Digital Humanities, 2020. Gent Gemapt, Retrieved 16 Augusts, 2021 from https://www.ghentcdh.ugent.be/projects/gent-gemapt

[22] Royal Library of Belgium, 2020, DATA-KBR-BE: Facilitating data-level access to KBR's digitized and born-digital collections for digital humanities research. Retrieved August 17, 2021 from https://www.kbr.be/en/projects/data-kbr-be/

[23] Philipp Schneider. 2021. Heraldry as a Historical Source to Conceptualize Medieval Spaces and Agents. *Data for History 2021: Modelling Time, Places, Agents* (2021). Retrieved July 10, 2021 from https://d4h2020.sciencesconf.org/data/pages/Schneider_Heraldry_as_a_Historical_Source_2.pdf

[24] Mapping Manuscript Migrations. 2021. Retrieved July 30, 2021 from https://mappingmanuscriptmigrations.org/en

[25] Agentschap Onroerend Erfgoed. 2021. Thesaurus. Retrieved August 18, 2021 from https://thesaurus.onroerenderfgoed.be/conceptschemes/ERFGOEDTYPES

[26] IIIF Consortium. 2021. IIIF Maps Technical Specification Group Charter, retrieved August 8, 2021 from https://iiif.io/community/groups/maps-tsg/charter/

[27] Bert Spaan. 2021. Allmaps & Annotorious. Retrieved August 18, 2021 from https://observablehq.com/@bertspaan/allmaps-and-annotorious?collection=@bertspaan/allmaps