nlGis: A Use Case in Linked Historic Geodata

Wouter Beek[1] and Richard Zijdeman[2,3]

[1] Dept. of Computer Science, VU University Amsterdam, NL
w.g.j.beek@vu.nl
[2] International Institute for Social History (IISH), Amsterdam, NL
richard.zijdeman@iisg.nl
[3] Faculty of Social Sciences, University of Stirling

Abstract. While existing Linked Datasets provide detailed representations of Cultural Heritage objects, the locations where the objects originate from is often not accurately represented. Countries, municipalities, and excavation sites are commonly represented by geospatial points, and the fact that countries and municipalities change their geometry over time is not reflected in the data.
We present *nlGis*, a collection of existing geo-historic datasets that are now published as Linked Open Data. The datasets in *nlGis* contain detailed geographic information about historic regions, with an emphasis on the Netherlands. We describe the creation of this Linked Geodataset and how it can be used to enrich Cultural Heritage data. We also distill several 'lessons learned' that can guide future attempts at publishing detailed Linked Geodata in the Cultural Heritage domain.

Keywords: Geodata, Linked Data, Cultural Heritage, GeoSPARQL, GIS

1 Introduction

Linked Open Data (LOD) is a flexible data representation paradigm that combines the expressive graph-based RDF data model with a distributed publication approach. Because it is grounded in Knowledge Representation, Linked Data allows heterogeneous data sources to be described with semantic detail. Furthermore, instead of making wholesale copies of the data, detailed fragments of data can be retrieved from the source location where the data is curated and maintained.

Because of these properties, LOD is well-suited for the publication of datasets in the Cultural Heritage [4] and Humanities [6] domain. Indeed, an increasing number of Cultural Heritage datasets is being disseminated as Linked Open Data, and initiatives exist to increase the publication of Linked Open Datasets in the humanities even further [3].

While existing Linked Datasets provide detailed representations of Cultural Heritage objects, the *locations* where these objects originate from, are often not accurately represented. Even when places are assigned unique identifiers, e.g., from GeoNames, they often do not provide detailed geographic information. At

the same time, the geographic extent is one of the most important – if not *the* most important – aspect of a location. Finally, often the same identifier is used to denote the same location through time, even though the geographic extent of that location may have changed.

When we look at the state of the LOD Cloud, it is not so strange that geographic information in Linked Cultural Heritage Datasets is often of limited quality or detail. There are very few LOD resources readily available that provide such detailed historic geometries. In fact, even datasets that focus specifically on representing (historic) geographic locations, represent such locations with very little detail, almost exclusively resorting to singular points and/or very rough bounding boxes. Finally, not all datasets use standardized vocabularies in order to represent geographic information. This severely reduces interoperability, since it no longer allows geometric information to be meaningfully queried across datasets. At the same time, geographic knowledge forms an important component of the context of Cultural Heritage object, and the need for Geographic Information Systems (GIS) support in the humanities has been recognized [7].

This paper presents *nlGis*, a collection of existing geo-historic datasets that are now published as Linked Open Data. The datasets in *nlGis* contain detailed geographic information about historic regions, with an emphasis on the Netherlands. The datasets in *nlGis* allow, for instance, a historic event to be located in a municipality at the time at which the event occurred, even though the municipality may no longer exist, or may have changed its geographic extent over time. Since almost every cultural object is related to a geographic location, either for its creation, performance, or conception, the potential applicability for these Linked Geodatasets within the Cultural Heritage domain is enormous.

While, the datasets in *nlGis* only provide a first step towards the full coverage of geographic information in Linked Data, we believe that the here described approach can be used to inform the publication process for other (historic) geographic datasets in the LOD Cloud. We specifically identify the weaknesses of existing Linked Geographic Datasets, their lack of detail, lack of temporality, and the fact that they are often not standards-compliant. The three source datasets that serve as input for the *nlGis* dataset are very different from one another: syntactically, semantically, and topically. We have formulated recurring issues and observations as lessons learned for others to use.

The rest of this paper is structured as follows: the next section presents related work on historic Linked Geodata, as well as related work on Linked Geodata standardization. Section 3 describes our approach of creating, storing, and querying the *nlGis* datasets. To conclude, we formulate lessons learned in Section 4.

2 Related Work

Geographical depictions of past events and contexts are important for historical scholars. Knowles et al. [5] distinguish between three mainstream applications of GIS in the first decade of the 21st century: the study of the history of land use,

the visualization of changing landscapes and urbanization, the construction of infrastructures that provide historical GIS data for others to visualize. Today, a fourth application called 'deep maps' can be added to this list. Deep maps are maps that connect multiple layers of information (e.g. photographs or people's experiences related to a specific location).

Instances of these four applications of GIS are not always disseminated in an optimal way. Products of historical GIS have been provided as images, as content inside viewer software (e.g. on a CD-ROM accompanying a book), or as downloads (ShapeFiles and tables) to be processed by special-purpose software such as QGIS[4]. The first two, images and interactive viewers, do not allow researchers to process or evaluate the data, whereas ShapeFiles and tables can only be easily linked if common vocabularies are applied. Moreover, web infrastructures are sometimes unable to disseminate data in the long run, as the historical GIS of Belgium[5]) sadly exemplifies.

While many Linked Datasets exist that include geographic information, such information is generally not very detailed. For example, the geographic region of France in today's DBpedia is the point at 2.35 longitude and 48.86 latitude. The historic dimensions of geographies of countries and places are even less well recorded. Even though geographic datasets have occupied a prominent position in different renditions of the LOD Cloud picture [6], even the most popular and comprehensive Linked Geodatasets, e.g., GeoNames, only contain simple point geometries. Even Linked Historic Datasets that focus on geographic information specifically are not very detailed and/or do not follow open standards, which makes them difficult to reuse.

2.1 Linked Geo-historic datasets

Portable Antiquities Scheme The British Museum and National Museum Wales host the Portable Antiquities Scheme website, providing support for the registration of archaeological objects found (by the public) in England and Wales. It provides a database of more than one million objects and uses the Heritage Data vocabulary[7] to describe objects and periods. To communicate the positions of findings, it relies on the Ordnance Survey Linked Data, consisting of a Gazatteer, postcode centroids and administrative boundaries. Unfortunately, there is no temporal variation in the spatial descriptions. So while it is possible to query for a finding from a particular region and period, the geographical result would always be depicted using contemporary information. Moreover, entities such as rivers, roads and districts are not described by detailed shapes such as lines or polygons, but by points.

[4] `https://qgis.org`
[5] `http://hisgis.be`
[6] `http://lod-cloud.net/`
[7] `http://heritagedata.org`

Nomisma Nomisma[8] aims to provide a vocabulary of numismatic concepts. While not specifically engaged with describing locations, concepts such as `nmo:hasFindspot` are used to describe finding places of coins and hoards. Such spots are – especially in the case of archaeological excavations – unlikely to be mere points, yet they are represented as such in the data. The Nomisma documentation does highlight the possibility of describing such points as having an approximate value, exemplifying the use of "a single point in lieu of the boundaries of a region".

Periodo Periodo[9] is an ontology published by the Institute of Museum and Library services, and focuses on transposing qualitative descriptions of time into Linked Data. It specifically raises awareness for time-specific descriptions of entities, in our case: changing boundaries. Unfortunately the ontology is geographically limited to `periodo:spatialCoverageDescription` in order to capture qualitative descriptions of geographical spaces and `dct:spatial` to link periods to locations in gazettteers.

Pleiades Pleiades is an RDF dataset that describes ancient geographic places [8]. This is one of the few Linked Historic Datasets that contains polygon geometries of the locations it describes. Unfortunately, it does not use a standardized vocabulary. As a result, this dataset cannot be queried with GeoSPARQL, and standards-conforming Linked Data tools cannot recognized that it contains geodata. Finally, the polygons described by the Pleiades RDF are rough bounding boxes that consist of four coordinates.

2.2 Linked Geodata standards

WGS84 Geo Positioning The WGS84 Geo Positioning vocabulary was created in 2003 by the W3C Semantic Web Interest Group[10]. While this vocabulary only allows the description of 2D and 3D points in the WGS84 coordinate reference system, it is used by a large number of Linked Open Datasets today.

GeoSPARQL GeoSPARQL [2] is an extension to the standard Semantic Web query language SPARQL, and has been standardized by the Open Geospatial Consortium (OGC)[11]. It contains a vocabulary for expressing topological relationships and functions (e.g., `geo:sfWithin`, `geof:sfEquals`), and the ability to represent various geometries (e.g., polygons, lines) in either the Well-Known Text (WKT) or the Geographic Markup Language (GML) format. By default, GeoSPARQL geometries use the WGS84 coordinate reference system, but allows other coordinate reference systems to be described on a per-geometry basis.

[8] `http://nomisma.org`
[9] `http://perio.do/technical-overview/#spatial-extent`
[10] `https://www.w3.org/2003/01/geo/`
[11] `http://www.opengeospatial.org/standards/geosparql`

3 Approach

This section presents the approach taken in creating *nlGis*, describing the transformation from source datasets to RDF (Section 3.1), the representation of geographic properties (Section 3.2), and how the data is published online for other to reuse (Section 3.3).

3.1 Transformation to RDF

The source datasets that are used in *nlGis* are published as ESRI ShapeFiles, CSV tables, and GeoJSON. Since Open Source resources for parsing the proprietary ESRI ShapeFile format are limited, we first convert this format into the XML-based Geographic Markup Language (GML) using GDAL[12]. This means that we have a variety of input data formats that all have to be converted into RDF.

While many tools exist that aim to support the conversion from source data into RDF, such conversion tools come with non-trivial limitations that make them impractical to use. In the construction of *nlGis*, we have come across the following teo main limitations. Firstly, existing conversion tools are memory-based. This means that they load the entire source dataset into memory before performing the specified transformation. Since memory is the most costly hardware resource (at least ten times more expensive than disk), this induces an unnecessarily high cost for the transformation process. Moreover, the vast majority of data transformation tasks do not require the entire dataset. Instead, they can be formed at the level of individual statements (e.g., when converting a GeoJSON arrary of floating point values into a Well-Known Text literal), or at the level of individual records (e.g., when asserting a relationship between two geometries of the same object). Secondly, existing transformation tools do not support a wide enough variety of input formats. For the creation of *nlGis*, we already use XML, JSON, and CSV input formats. In the case of GeoJSON and GML, there are non-trivial extensions to the base languages – JSON and XML, respectively – that would ideally also be covered by transformation tools.

A full analysis of existing tools for data transformation, including a comparison of their respective benefits and shortcomings, is outside the scope of this paper. We therefore resort to custom scripts in order to create *nlGis*. While a process that is as open-ended as data transformation may in the end require a full programming language, we believe that the main contribution of transformation tools lies in the abolity to describe a transformation process in a way that can be shared with others who are using the same tools. The benefit is not necessarily ease of use or automation, but improved documentation and communication with others. RML [13] provides such a declarative representation format for expressing transformations, but RML tools do not yet meet the above stated requirements.

In addition to the datasets we transform, the Historic Dutch municipalities dataset[14] was already published as Linked Data. This has a tremendous benefit,

[12] `http://gdal.org`
[13] `http://rml.io`
[14] `http://gemeentegeschiedenis.nl`

since this dataset can be downloaded using the Follow Your Nose principle. According to this principle, it is possible to start at some online location within the dataset, and traverse the entire online graph in order to obtain all information.

3.2 Geographic representation

In the conversion to RDF, we specifically want to focus on how geographic data is best represented. The two most popular vocabularies for representing geographic Linked Data are the WGS84 Geo Positioning vocabulary and GeoSPARQL (Section 2.2. We perform measurements on a very large (>650K documents, >38B triples), but also somewhat outdated (late 2015), LOD Cloud scrape performed by the LOD Laundromat[15]. Unfortunately, there is not a more recent scrape of comparable size to perform a more up-to-date analysis on.

Within this scrape, the WGS84 Geo Positioning vocabulary is used in many more documents (11,235) than GeoSPARQL (47). Table 1 gives an overview of the use of the various properties. While individual longitude (`wgs85:long`) and individual latitude (`wgs84:lat`) are both asserted approximately 43M times in 11K documents, there are 33,422 more assertions of the former. Since there are very few cases where asserting a longitude without a latitude makes sense, this may indicate a geo-specific data quality issue. While property `wgs84:lat_long` has modeling benefits over the use of individual longitude and latitude properties, it is almost never used (283 statements; 173 documents). The altitude property (`wgs84:alt`) is used relatively frequently (2.3M triples; 9.8K documents), especially given the fact that Linked Geodata is not often visualized on a map that is able to display altitude.

Even though GeoSPARQL is used in only 47 documents, these documents contain relatively many GeoSPARQL assertions. In fact, overall there are more GeoSPARQL (188M) than WGS84 Geo Positioning (42M) geometries, even though they appear in a relatively tiny amount of documents. Most GeoSPARQL geometries are points (165,875,711 statements, or 88%), 6% are polygons, and the remaining 6% are linestrings. All geometries are serialized in Well-Known Text (WKT), and none are serialized in GML. Notice that it is possible for `geo:asGML` to never be used, but still appear in one document: it appears in the GeoSPARQL vocabulary itself.

3.3 Data publication

An overview of the *nlGis* datasets is given in Table 2. CShapes [9] encodes detailed historical maps of state boundaries and capital cities from the second World War onward. Countries are coded according to the Correlates of War and the Gleditsch & Ward state lists. The RDF version of CShapes contains information about 207 countries and 201 capital cities. The historic Dutch municipalities dataset[16] contains 1,679 historic municipalities, including the relationships between them,

[15] `http://lodlaundromat.org`
[16] `http://gemeentegeschiedenis.nl`

Table 1: Overview of the use of standardized vocabularies, based on the LOD Laundromat scrape. Usage is quantified in terms of (i) the number of documents and (ii) the number of triples, in which the respective properties appear.

Property	№ statements	№ documents
wgs84:alt	2,349,607	9,843
wgs84:lat	42,883,363	11,134
wgs84:lat_long	283	173
wgs84:location	14,688,561	117
wgs84:long	42,916,785	11,134
geo:asGML	0	1
geo:asWKT	188,427,329	50
geo:hasGeometry	28,366,268	7

Table 2: Statistical overview of the *nlGis* datasets.

Dataset	№ statements	Main concepts	№ geometries
CShapes	6,120	countries, cities	510
Mint Authorities	6,987	authorities, houses	950
Gemeentegeschiedenis	46,929	municipalities, provinces	3,219
Total	60,036	features, geometries	4,679

e.g., two or municipalities are often merged into one. The Mint Authorities of the Low Countries[17] contains the polygons of the major coin issuing authorities that existed in the Low Countries in the Middle Ages. Each authority is paired with begin and end dates. Starting from the twelfth century onward, most authorities are included, except for small authorities such as towns.

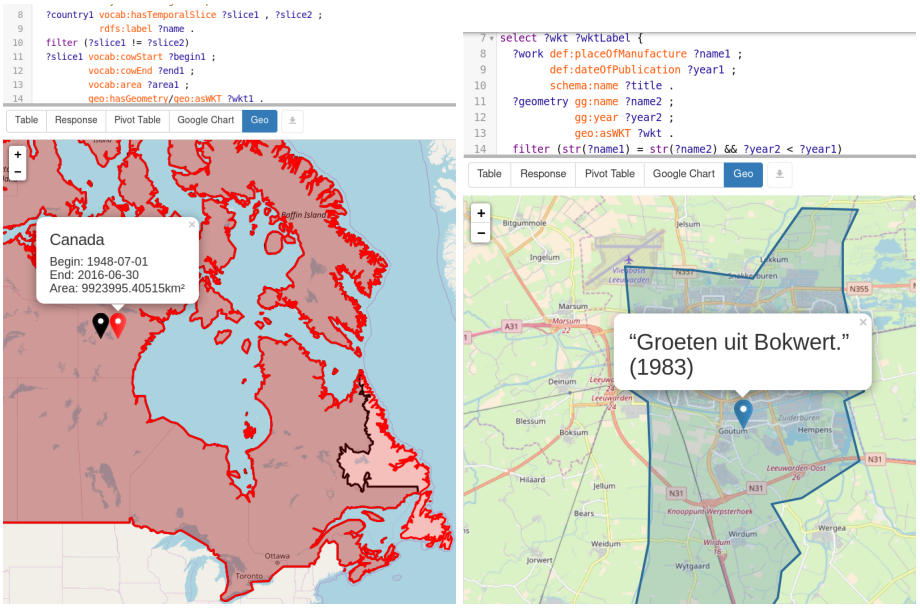
The *nlGis* datasets are stored on the Druid Linked Data platform[18], where the data can be browsed and queried online. Druid is developed within the CLARIAH project[19], and is hosted by the Royal Dutch Academy of Arts and Sciences (KNAW). Datasets are stored using Header Dictionary Triples (HDT)[20], which allows them to be stored on disk rather that in memory (which incurs a relatively low hardware cost).

The writing of complex queries is an iterative process, in which inspection of the results for the previous query inform the (re)writing of the next query. For querying Linked Geodata, it is important to use a GeoSPARQL-compliant editor like GeoYASGUI [1] (Figure 1).

Figure 1a gives an example of how the data in *nlGis* can be used in practice. This example shows how the geospatial extent of Canada changes over time. Figure 1b shows how *nlGis* can be used to enrich Cultural Heritage data. This query is written over the knowledge graph of the International Institute for

[17] https://datasets.socialhistory.org/dataverse/lowcountries_GIS
[18] https://druid.datalegend.net/nlgis
[19] https://clariah.nl
[20] http://www.rdfhdt.org/

Fig. 1: The results of two GeoSPARQL queries over *nlGis* data. Results are displayed in the GeoYASGUI result set viewer.



(a) Retrieves the change in Canada's geospatial extent: Newfoundland and Labrador were added to Canada in 1948.

(b) Retrieves the geospatial extent of locations in the Netherlands where cultural heritage objects are published.

Social History (IISH), which records information about hundreds of thousands of cultural history objects. The geospatial extents from *nlGis* are queried through SPARQL federation, which allows data that is published in a distributed way to be integrated on a per query basis. In fact, this query specifically retrieves the geo-temporal extent that coincides with the date of publication of the particular cultural heritage object.

4 Leassons Learned & Conclusion

We conclude by identifying the main lessons we have learned during the creation, publication, and use of *nlGis*. The following lessons learned can be used in order to ease future publications of Linked Geodata in the Cultural Heritage domain:

Combine components that belong together Source data often stores components of values that conceptually belong together in separate properties. For example, CShapes stores the begin and end dates of geo-temporal extensions in one, two, or three properties (depending on their availability: the year, month, and day property). Another example is the separate storage of longitude and latitude properties that together represent a point geometry.

It is better, e.g., more fault tolerant, to represent values that conceptually belong together with one property. *nlGis* stores the begin and end dates of geo-temporal extents in one property, whose value is of type `xsd:gYear`, `xsd:gYearMonth`, or `xsd:date`, depending on how many date components are specified. No functionality is lost, since SPARQL contains functions that return the constituent components of dates (`year()`, `month()`, and `day()`). The longitude and latitude of a point geometry can are also stored together in *nlGis*, using one WKT literal. In some cases, combining longitude and latitude solves an important data quality issue: if a location has more than one point geometry, it is no longer clear which `wgs84:long` and `wgs84:lat` values belong together.

Do not use ambiguous default values Default values are very popular in non-RDF sources. Default values are most commonly represented by values that are (assumed to be) nonsensical within a certain context. For example, in CShapes the year -1 denotes the fact that a year is not known. This convention makes some sense in the context of CShapes, which only contains geographic locations after 1945. However, in Linked Data we cannot rely on such dataset-specific disambiguation assumptions. For example, the LOD Cloud may well contain geometries whose begin date is the year -1. This is why we try to detect and remove ambiguous default values within the transformation process. If a property has a default value in the source data, we simply do not generate a triple for that particular property in the RDF data.

No perfect tool for data transformation It is not easy to find a data transformation tool that 'does the job' (see Section 3.1.

No perfect triple store for GeoSPARQL We tried out three production-grade triple stores that claim to support GeoSPARQL, but did not find one that is able to do so correctly and with good performance. Table 3 enumerates our findings in terms of (i) standards-compliance, (ii) correctness, and (iii) performance. As with data transformation tools, a detailed comparison of GeoSPARQL support in different triple stores is outside the scope of this paper. Even through GeoSPARQL support is not yet perfect, it certainly is usable, and the here mentioned three triple stores are specifically working on improving the support over time.

Direct geospatial feedback Use a SPARQL editor that is able to detect and display geospatial data for direct feedback.

Interoperable representation In order to make geospatial data inter-operable with others, and directly useful in standards-compliant tools, it is best to use widely supported and standardized representations. There is no reason to the use the WGS84 Geo Positioning vocabulary anymore: WGS84 point geometries can also be expressed in the standardized GeoSPARQL vocabulary. In GeoSPARQL, use WKT in order to serialize geometries, since GML is (almost) never used (see Section 3.2).

Table 3: Experience-based comparison of GeoSPARQL support in three production-grade triple stores.

Triple store	**Standards-compliance**	**Correctness**	**Performance**
Virtuoso	Uses custom relations and functions	Uses bounding boxes rather than the geometries themselves	Can be used interactively and/or to power an application
GraphDB	Compliant	Once a result is obtained it seems to be correct	Cannot be used interactively; many queries receive a timeout
Stardog	Only few functions are implemented; some are not part of GeoSPARQL	Could not test	Geographic queries with polygons do not terminate.

References

1. W Beek, E Folmer, L Rietvel, and J Walker. Geoyasgui: The geosparql query editor and result set visualizer. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42, 2017.
2. Open Geospatial Consortium et al. Ogc geosparql-a geographic query language for rdf data, 2012.
3. Rinke Hoekstra, Albert Meroño-Peñuela, Auke Rijpma, Richard Zijdeman, Ashkan Ashkpour, Kathrin Dentler, Ivo Zandhuis, and Laurens Rietveld. The datalegend ecosystem for historical statistics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2018.
4. Eero Hyvönen. Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1):1–159, 2012.
5. Anne Kelly Knowles and Amy Hillier. *Placing history: How maps, spatial data and GIS are changing historical scholarship*. ESRI, 2015.
6. Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. Semantic technologies for historical research: A survey. *Semantic Web*, 6(6):539–564, 2015.
7. J.B. Owens, May Yuan, Monica Wachowicz, Vitit Kantabutra, Emery A. Coppola Jr., Daniel P. Ames, and Aldo Gangemi. Visualizing historical narratives: Geographically-integrated history and dynamics gis. 2009.
8. Rainer Simon, Leif Isaksen, Elton Barker, and Pau de Soto Cañamares. The Pleiades gazetteer and the Pelagios project. *Placing Names: Enriching and Integrating Gazetteers*, 2015.
9. Nils B. Weidmann, Doreen Kuse, and Kristian Skrede Gleditsch. The geography of the international system: The cshapes dataset. *International Interactions*, 36(1):86–106, 2010.