# The Lure and Limits of Linked Data: the case of World Historical Gazetteer

Karl Grossner and Ruth Mostern
University of Pittsburgh, US

## Introduction

Spatial humanists and others have long recognized the enormous integrative potential of using places as common points of reference for heterogeneous information. To realize that potential, collections of named places must be *abundant*, *diverse*, *collectively assembled*, and *historically deep*. In 2017, the World Historical Gazetteer (WHG) project based at the University of Pittsburgh undertook to build a freely available web platform[1] that would facilitate the collaborative development of such a collection, and to provide multiple ways of accessing its continuously growing results. Version 1 of the WHG platform was initially launched in 2019, and in 2021 voted "Best DH Tool or Suite of Tools" in the annual Digital Humanities Awards. Version 2 followed in April 2022, and a Version 3 is due in mid-2024.

The approach taken by the WHG project for assembling, linking, and publishing diverse place data as a free web resource utilizes the technological and social elements of the Linked Data paradigm (LD), as its characteristics match the requirements of a comprehensive digital historical gazetteer well (Bol 2011, Grossner, Keßler, and Janowicz 2014). These include (i) *extensibility*, due to its underlying graph-based conceptual model; (ii) *multivocality*, by accommodating multiple possibly conflicting statements about the same phenomena; (iii) *integration* and (iv) *sustainability*--both facilitated by an expressive standard interchange format expressed in RDF[2].

The union index has grown to well over 2 million "clusters" of place attestations, i.e. sets of attestations for the same (or closely matched) place from multiple sources. Many additional datasets are published in the WHG and not yet added to the union index, and many more are at some earlier stage of accessioning. The WHG is in fact *collectively assembled*, and is well on its way to being *abundant*, *diverse*, and *historically deep*.

## A different kind of gazetteer

The WHG is not so much a gazetteer as it is a collection of gazetteers, generically termed *place datasets* in the platform. The records from datasets published in the WHG are to a large extent internally linked by their creators in a "union index," and accessed via faceted search and an application programming Interface (API). Individual datasets are also presented as publications within the system and can be browsed and queried as such. The WHG platform provides features for performing the linking of data and disseminating the results as truly "linked open *usable* data" as advocated and described by Sanderson (2020).

## The lure and the promise

Knowledge about the past derived from research outputs, archives, and library holdings can be brought together indirectly with Linked Data methodology by common references to places. In Figure 1, each project (clear circle) has some information pertaining to Tbilisi, concerning perhaps museum holdings or historical events. Each project has within its research output a listing of all the places its work references--including Tbilisi. For each place they have identified one or more identifiers from an

[1] World Historical Gazetteer, http://whgazetteer.org

[2] RDF (Resource Description Framework) is the core *subject-predicate-object* data model underlying Semantic Web and Linked Data methodologies

"authority" resource such as Getty Thesaurus of Geographic Names (TGN), Bibliothèque Nationale de France (BnF), GeoNames, or Wikidata (green circle). By publishing their place records in the WHG, projects are in effect announcing "we have information about {x} and Tbilisi." A search for "Tbilisi" -- or any of the 70 name variants gathered from linked records -- will currently return a set of 7 attestations, each from a different source.

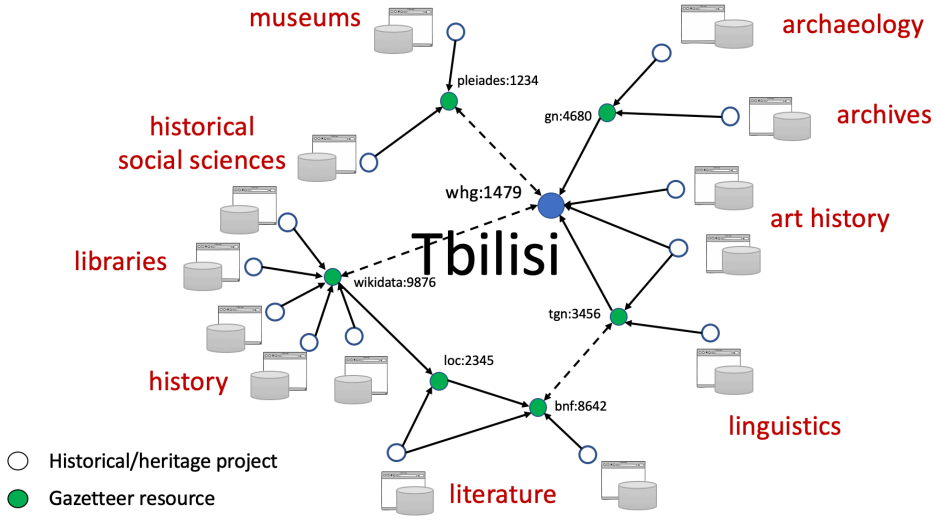


*Figure 1 - Linking projects and disciplines with place*

*Multivocality*. Linked Data methodologies facilitate the surfacing of suppressed place names and difficult histories by supporting peoples' discoveries about past places. It can allow genealogists and others to discover common historical connections to places, even if ancestors had different experiences at them and may have called them by different names. A visitor to the WHG who searches for *Ayers Rock* finds information about *Uluru*. A search for *Tenochtitlan* returns *Mexico City* and *Ciudad de México* (and vice versa), and a search for *Batavia* links to *Jakarta* as well.

*Teaching*. An index of linked gazetteers is a powerful teaching tool. By exploring how the same name recurs across the globe, students can trace contours of immigration, conquest, and political power. The WHG *Place Collection* and *Collection Group* features support classroom exercises for creating and annotating collections of thematically linked places.

## The limits

*Sparse temporal information*. Relatively few historical place attestations include timespans indicating a period of existence; publication year of the source is often all that is available. For this reason, it is not possible to get comprehensive results when filtering on a year, timespan, or period, or to make clear the distinction.

*LD is not (necessarily) curated*. A stated premise of the original RDF model design was that "anyone can say anything about anything" (W3C 2002). This is a blessing and a curse: it affords essential multivocality, but the quality of an information resource can suffer, and contributors to an LD graph have no control over who says what about their statements.

*Disambiguation and conflation*. The requirement for one record per place is a burden for many potential collaborators. Places can have multiple names, types, extents/locations and relationships over time. Aggregating these attributes within a single record can be difficult.

*Semantics*. There is little agreement as to some essential categorizations, e.g. of place type. The WHG allows any term to be added for type, but because contributors resist mapping their terms to the common vocabulary we offer, reconciliation results and filtering of search results by place type are somewhat hampered.

## Looking Forward

Historian Jo Guldi recently asked how to take a digital, quantitative approach to history that still maintains the complexity of past human experience and the heterogenous, ambiguous, and ideologically embedded sources in which it is represented (Guldi 2023). Geographer Ruth Wilson Gilmore argues that struggles for racial justice are always also struggles for place (Gilmore 2022). Linking multiple digital humanities projects together is on its face a worthwhile goal, but there is still work to be done to determine how best and most ethically to do that while honoring the fact that each project has its own unique and organic relationship with a data-sharing community, one that may be vulnerable and may have a history of exploitation (Smith 1999). This is a complex practical and epistemological challenge, one that linked data makes both easier and more complex in various ways, and with which the WHG continues to wrestle.

## References


Bol, P. K. (2011). What Do Humanists Want? What Do Humanists Need? What Might Humanists Get? In *GeoHumanities*, edited by M. Dear, J. Ketchum, S. Luria, and D. Richardson, 296–308. Oxford, UK: Routledge.

Gilmore, Ruth Wilson (2022), "Fatal Couplings of Power and Difference: Notes on Racism and Geography" in Ruth Wilson Gilmore, Brenna Bhanda and Alberto Toscano, *Abolition Geography* (Verso).

Grossner, K., Keßler, C. and Janowicz, K. (2016). Place, Period, and Setting for Linked Data Gazetteers. In R. Mostern, H. Southall, and M.L. Berman (Eds.). *Placing Names: Enriching and Integrating Gazetteers*. Bloomington: Indiana University Press.

Guldi, Jo (2023), *The Dangerous Art of Text Mining: A Methodology for Digital History* (Cambridge)

Sanderson, R. (2020). *LOUD*. Web page, retrieved on 09 Feb 2024 from https://linked.art/loud/.

Smith, Linda Tuwahi (1999), *Decolonizing Methodologies: Research and Indigenous Peoples* (Bloomsbury)

W3C (2002). *Resource Description Framework (RDF): Concepts and Abstract Data Model*. retrieved on 14 Feb 2024 from http://www.w3.org/TR/2002/WD-rdf-concepts-20020829/