Computers, Environment and Urban Systems 64 (2017) 309–320

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A quantitative analysis of global gazetteers: Patterns of coverage for common feature types

Elise Acheson [a], Stefano De Sabbata [b], Ross S. Purves [a]

[a] University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[b] Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom

ARTICLE INFO

Article history:
Received 21 September 2016
Received in revised form 28 January 2017
Accepted 13 March 2017
Available online 13 April 2017

Keywords:
Gazetteers
Data quality
GeoNames
Placenames
Geocoding

ABSTRACT

Gazetteers are important tools used in a wide variety of workflows that depend on linking natural language text to geographical space. The spatial properties of these data sources, such as coverage, balance, and completeness, affect the performance of common tasks such as geoparsing and geocoding. However, little attention has focused on how these properties vary in global gazetteers, particularly across country boundaries and according to feature types. In this paper, we present a detailed investigation of the spatial properties of two open gazetteers with worldwide coverage: GeoNames, and the Getty Thesaurus of Geographic Names (TGN). Using point density maps, correlations, and linear regressions, we analyze the global spatial coverage of each data source for the full set of features and for top feature types: populated places, streams, mountains, and hills. Results show wide discrepancies in coverage between the two datasets, sharp changes in feature type coverage across country borders, and idiosyncratic patterns dominated by a few countries for the more sparsely covered natural features. As more and more systems rely on recognizing and grounding named places, these patterns can influence the analysis of growing amounts of online text content and reinforce or amplify existing inequalities.

1. Introduction

Gazetteers play a central role in linking text to space, influencing a multitude of application outcomes through their use in tasks such as identifying placenames[1] in text, disambiguating placename references, and associating placenames with a geographical footprint and type information. Until recently, gazetteers were primarily produced top-down, typically as curated resources for placenames in a prescribed area such as a country. Today, with data easily stored and shared online, and vast quantities of data released as open data, the ways in which gazetteers are being produced and distributed is evolving. At one end of the spectrum remains a top-down, strongly regulated process, where organizations such as national mapping agencies produce gazetteers according to explicitly defined data quality standards and local laws. At the other end are crowdsourcing efforts collecting information about places from anyone who wishes to contribute, often largely relying on the notion of the 'wisdom of the crowd' principles for data quality (Goodchild & Li, 2012). Somewhere on this spectrum are two gazetteers with some level of data curation, nominally global coverage, but limited explicit information with respect to data quality: GeoNames (GeoNames, 2016), and the Getty Thesaurus of Geographic Names (TGN, 2016).

These gazetteers form the focus of the present paper. Perhaps because of their worldwide coverage and their ready availability, both are popular in many research projects and applications, with GeoNames arguably the most commonly used gazetteer today. Despite this popularity, there has been limited scrutiny of its contents, with attention typically limited to a particular region or country, and focused largely on populated place features rather than a broader set of feature types. Smart et al. (2010) mapped the overall coverage of GeoNames in Great Britain, contrasting it with national mapping agency data and crowdsourced datasets. Ahlers (2013) conducted a broader examination of data quality in GeoNames, identifying anomalies and quality indicators for populated places in Central America, Germany, and Norway. Looking at both GeoNames and TGN, De Sabbata and Acheson (2016) quantitatively compared their coverage for all features and populated places in Great Britain, finding the datasets less detailed and less balanced than national mapping agency data. Although these studies have revealed that coverage in these products is unbalanced even within individual countries, the overall picture remains unclear since to date, an in-depth systematic global analysis, looking across country boundaries and at a range of feature types across gazetteers, has not been carried out.

An initial exploration of such properties examined global coverage of GeoNames alone and explored the distribution of a single feature type

E-mail addresses: elise.acheson@geo.uzh.ch (E. Acheson), s.desabbata@leicester.ac.uk (S. De Sabbata), ross.purves@geo.uzh.ch (R.S. Purves).

[1] We use the more vernacular term *placename* interchangeably with *toponym* in this paper.

http://dx.doi.org/10.1016/j.compenvurbsys.2017.03.007

(populated places) as a function of population (Graham & De Sabbata, 2015). Expanding on this work, we undertake a detailed comparative investigation of the global spatial properties of both GeoNames and TGN. We not only look at the full datasets, but also present a worldwide analysis of coverage for the four most frequent feature types in GeoNames, matched with corresponding types in TGN: populated places, streams, mountains, and hills. These four feature types account for a large portion of the full datasets in both gazetteers, particularly populated places which comprise over a third of all the data in both GeoNames and TGN. As for streams, mountains, and hills, they are among the most common natural features found in the data sets, and in the case of mountains, the most commonly referenced examples of a geographic feature in empirical experiments (Smith & Mark, 2001). Understanding the global coverage of these named natural features is particularly important in the context of any work analyzing the distribution of common toponym types (Campbell, 1991) and analysis of texts containing references to natural features (Moncla et al., 2014). For both gazetteers, we examine and compare feature distributions at fine, medium, and coarse granularities.

As discussed in the review that follows, *coverage* and *balance* are two pivotal quality indicators to assess the fitness for use of gazetteers for many common tasks. We therefore pose the following research questions:

1. How do GeoNames and TGN compare in terms of overall global coverage and balance?

2. How are important feature types in GeoNames and TGN distributed globally, and how do they compare in terms of coverage and balance?

We review previous work focusing on gazetteer properties, sources, and quality, as well as tasks in which gazetteers play a role. We then introduce in more detail the properties of the two gazetteers we analyzed, before setting out the analysis methods to characterize and compare GeoNames and TGN. Our results are presented as both graphical and numerical data, before we discuss their implications, particularly in terms of the suitability of these data sources for relevant tasks. We conclude with a list of key gazetteer shortcomings and propose future research focused on addressing these.

2. Gazetteers

> "*There is remarkable diversity in approaches to the description of geographic places (…)*".
>
> [Linda Hill, Georeferencing, p. 94]

Gazetteers are resources that store structured information about places, minimally providing name, type, and location (or footprint) information for each place or record (Hill, 2000; Mostern et al., 2016). Each record may also contain other attributes such as alternative names, population information for populated places, and containment relationships - for example which country or region the place is in. Records may contain links to matching records in other datasets. These 'linked data' records are ones deemed to be about the same place through a matching process that, for instance, compares text, positional, and type information across resources (Sehgal et al., 2006; Smart et al., 2010). Placenames have in fact become a central node in linked open data, with GeoNames lying at the center of the linked open data cloud diagram (Schmachtenberg et al., 2014), demonstrating the efficacy of placenames as a way of relating information in the developing semantic web.

2.1. Gazetteer sources and production

Gazetteers have traditionally been produced in a top-down process, most commonly by national mapping agencies to serve as official placename resources for a defined area of interest such as a country, sometimes under specific legal or regulatory conditions. For example, Ordnance Survey (OS) produces the OS 1:50k gazetteer (2016) (and more recently, OS Open Names) for the extent of Great Britain, and SwissTopo produces SwissNames 3D (2016) for the extent of Switzerland. In the case of the United States, examples include a national resource for domestic names, the Geographic Names Information System (GNIS, 2016), developed by the U.S. Geological Survey, and an international resource for foreign names, the GEOnet Names Server (GNS, 2016), developed by the National Geospatial-Intelligence Agency.

As well as general purpose gazetteers, typically created by national mapping agencies and other government authorities, purpose-built gazetteers are created for a wide range of purposes. Among these are the TGN, a structured gazetteer with the aim of improving access to art, architecture, and material culture by enabling indexing. Due to its focus on these topics, historical names are important elements of the TGN, allowing links of historical artifacts to be made between present day locations and texts describing them in a historical context.

More recently, gazetteers have also been produced by incorporating bottom-up methodologies, where data is collected from multiple sources and integrated. Two heavily used global spatial datasets, OpenStreetMap and GeoNames, are produced this way: their sources include authoritative data, such as those described above where licensing permits, but also original data contributed by individuals, also known as volunteered geographic information (VGI) (Goodchild, 2007). Further still along the spectrum from top-down to bottom-up production are approaches to creating structured gazetteers using only crowdsourced data, through the extraction, analysis, and merging of multiple sources. One such example, the Gazetiki project, mined Wikipedia and Panoramio data to automatically create a gazetteer, relying on linguistic cues, search hits, and the GeoNames feature type hierarchy for entity typing (Popescu et al., 2008).

A complementary body of research focuses on both augmenting and enriching existing gazetteers and the generation of so-called meta-gazetteers to build better resources, whether more complete (with more features, or with richer annotation for existing features), or deemed more suitable for a particular task (Kessler et al., 2009; Smart et al., 2010). In one example using VGI, Gao et al. (2017) present a framework for efficiently creating new gazetteer entries from large numbers of user-tagged photographs, many of which contain feature types like 'park', 'museum', or 'river' as tags. Finally, OpenStreetMap has also been used as a gazetteer source directly, or to augment existing placename resources (de Oliveira et al., 2016; Hess et al., 2014; Yin et al., 2014).

As feature types are one of the three basic requirements of a gazetteer entry (Hill, 2000), any work seeking to integrate or augment gazetteers faces the challenge of assigning appropriate types to features, and potentially having to align different feature type ontologies to each other. A common use case in gazetteer conflation is to consider feature type information as evidence of (dis)similarity when trying to detect whether records are about the same feature (Fu et al., 2005; Hastings, 2008; Smart et al., 2010). However, this is a challenging task since feature types may vary widely between gazetteers, and the process of feature type alignment is itself complex (Janowicz & Keßler, 2008; Zhu et al., 2016). These difficulties are illustrated by for example Fu et al. (2005) who established "equivalence links" between feature type hierarchies, but found that strong constraints on feature type alignment led to poor performance. The underlying problem is further illustrated by Smart et al. (2010) who noted that even in national mapping agency data, large proportions of features were simply classified as "other". Zhu et al. (2016) recognize this challenge and combine top-down ontology analysis with bottom-up data-driven methods using spatial signatures related to instances of feature types to explore alignment issues in GeoNames, TGN and DBPedia Places.