Future Generation Computer Systems 54 (2016) 389–398

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data

Silvio D. Cardoso[a],*, Flor K. Amanqui[a], Kleberson J.A. Serique[a], José L.C. dos Santos[b], Dilvan A. Moreira[a]

[a] *Department of Computer Science, University of São Paulo, São Carlos, Brazil*

[b] *National Institute for Amazonian Research, Manaus, Brazil*

HIGHLIGHTS

- This work provides an architecture to Semantic and Volunteered Gazetteers.
- Improves geographical coordinates from biodiversity repositories.
- Shows a prototype from a Semantic Gazetteer.

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form
1 May 2015
Accepted 6 May 2015
Available online 16 May 2015

Keywords:
Semantic Web
Volunteered geographic information
Gazetteer

ABSTRACT

Current implementations of gazetteers, geographic directories that associate place names to geographic coordinates, cannot use semantics to answer complex queries (most gazetteers are just thesauri of place names), use domain ontologies for place name disambiguation, make their data sets available in the Semantic Web or support the use of Volunteered Geographic Information (VGI). A new generation of gazetteers has to tackle these problems. In this paper, we present a new architecture for gazetteers that uses VGI and Semantic Web tools, such as ontologies and Linked Open Data to overcome these limitations. We also present a gazetteer, the Semantic Web Interactive Gazetteer (SWI), implemented using this architecture, and show that it can be used to add absent geographic coordinates to biodiversity records. In our tests, we use this gazetteer to correct geographic data from a big sample (around 142,000 occurrence records of Amazonian specimens) from SpeciesLink, a big repository of biodiversity collection records from Brazil. The tests showed that the SWI Gazetteer was able to add geographic coordinates to around 30,000 records, increasing the records with coordinates from 30.29% to 57.5% of the total number of records in the sample (representing an increase of 90%).

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Using information about biodiversity data from the literature [1], it is possible to highlight the main challenges faced when analyzing this kind of data: (i) deal with large volumes of information, (ii) achieve interoperability of information from different sources and formats, (iii) manipulate data and images, (iv) handle geographic information. Data from biological repositories have a large number of records with inaccurate geographic information [2]. The lack of precise geographic spatial information in biological records can lead to problems, such as the precise demarcation of protection areas for endangered species [3].

Geographic Information Retrieval (GIR) techniques can be used to improve the accuracy of geographic information, contained in biological data, and store it in a gazetteer. A gazetteer is a geographic directory that associates place names to geographic coordinates. They are commonly implemented as directories that contain triples of place names (N), feature types (T) for named geographic places, and geographic footprints (F) with geographic coordinates [4,5]. Gazetteers are important for allowing geospatial queries, such as "find all rivers in Jaú National Park", to be performed by GIR systems [6].

While most current gazetteers are held by small and well-defined groups, the next generation of gazetteers needs to incorporate updates from Volunteered Geographic Information (VGI) [7]. The term VGI is relatively new and strongly adopted

* Corresponding author. Tel.: +55 16 3373 9997.
E-mail addresses: silvio.domingos.cardoso@usp.br (S.D. Cardoso), flork@icmc.usp.br (F.K. Amanqui), serique@icmc.usp.br (K.J.A. Serique), lcampos@inpa.gov.br (J.L.C. dos Santos), dilvan@icmc.usp.br (D.A. Moreira).

http://dx.doi.org/10.1016/j.future.2015.05.006
0167-739X/© 2015 Elsevier B.V. All rights reserved.

to describe content generation by users on the web. However, in the biodiversity community, VGI has had a long tradition and important role, because it enables the monitoring of species, including endangered ones, providing a substantial support to researchers. Moreover, the use of VGI can simplify the data storage and management in gazetteers, especially if manual intervention is required [8].

Kessler et al. [6] discuss the recent challenges in gazetteer research and point out missing pieces required for the next generation of gazetteer infrastructure. They point out problems like the lack of: semantics to answer complex queries (most gazetteers are just thesauri of place names), infrastructures to support Linked Data, and domain ontologies for gazetteer construction and place name disambiguation. They do not mention any support for VGI.

The term Linked Data refers to a set of best practices for publishing and connecting structured data over the Web [9]. Linked Open Data (LOD) refers to the use of Linked Data with open data. It has been adopted by an increasing number of data providers, leading to the creation of a global open data space containing billions of assertions—the Web of Data or Semantic Web [9]. These assertions include data about almost everything, from movie names to biochemical reactions. The SPARQL language[1] (a query language for the Resource Description Framework—RDF) is used to access data repositories in the Semantic Web (called SPARQL Endpoints) [10]. RDF is a directed, labeled graph data format for representing information in the Semantic Web. SPARQL is similar in syntax to SQL and can find data based on their relationships with other data. In the case of SPARQL however, these relationships can be semantic, and the search can be considered a semantic search. For instance, in SPARQL it is possible to ask for records that have implicit information, e.g., ask for specimens of the order Anura (frogs) and get records with data where the strings "Anura" or "frog" do not appear.

Currently, some well known gazetteers incorporate the use of VGI, for example, WikiMapia and GeoNet. However, their data (and voluntary information) are not available in the Semantic Web. The next generation of gazetteers needs to boost VGI use, together with LOD, to improve information update and completeness.

The new generation of gazetteers [6] focuses on the integration of VGI, semantics-based retrieval and navigation. Among the most important points they propose for this new generation, we highlight:

- The shift in focus towards non-expert users and machine-to-machine communication (e.g., for reasoning and harvesting).
- Extraction of sites from established data sets (such as government agencies) and from implicit geographic information.
- Data trust based on authority (i.e., data providers) and on trust models as proxies for VGI quality.
- Data provenance ensured by authorities and inferred from trust ratings and history of user interactions. In the case of biodiversity data, it can be hard to find official names for wild places.
- Ontology based inference of additional facts about features and feature types. This can help to overcome the shortcomings of natural language descriptions used in existing gazetteers.

In this paper, we present a new gazetteer architecture that fulfills the requirements for this new generation. Our aim is to tackle the challenges presented by the authors in [6], such as support to LOD and VGI principles [9]. We also present the Semantic Web Interactive Gazetteer (SWI), a prototype implementation of this architecture. To validate it, we used a data set from SpeciesLink,[2] a large repository about important Brazilian biodiversity records. We chose this repository because it holds important data about Amazonian biodiversity but has a large number of records with inaccurate geographic information [11].

Using SWI, it was possible to find geographic coordinates for around 57% of records from SpeciesLink that did not have them before. These added coordinates were then manually verified by sampling them.

The remainder of this article proceeds as follows: Section 2 discusses related works. Sections 3 and 4 show the architecture and prototype implementation. Section 5 presents the data set used to validate our work and Section 6 discusses the experiments and their results. Finally, in Section 7, we conclude this work summarizing our results and describing future works.

2. Related work

We conducted a systematic review in the literature for works that dealt with VGI or semantic-based retrieval, which are the focus of SWI. These two techniques together with semantic-based navigation are the focus of the agenda for the next generation gazetteer infrastructure [6] (proposed after a review of recent challenges in gazetteer research). Despite that, we only found few works. The articles found use just VGI or semantic retrieval, but not both. They use thesauri (dictionaries that contain place names) or ontologies to support place name disambiguation. Below we list them:

- The DIGMAP Gazetteer [12] enables the search for historical places using a Web Service, which receives queries and returns results in XML format. The Semantic information, used in the DIGMAP Gazetteer to disambiguate name places, comes from the use of the GeoNetPT ontology. However, the gazetteer's data are not available in the Web of Data (as Linked Data) because they are not accessible using a SPARQL endpoint. Furthermore the DIGMAP Gazetteer does not support the use of VIG.
- The KIDGS Gazetteer [13] enables users to search place names in precise or imprecise ways, e.g., "north of Beijing", "airport in Beijing", "Beijing". Users can make queries, using XML files, in the format: subject, predicate, object. These queries are sent to a Web Service that processes them using KIDGS Gazetteer. KIDGS performs place disambiguation using ontologies. However, KIDGS neither supports VIG nor makes its data accessible on the Semantic Web (data are stored in an internal relational database).
- The FODGS Gazetteer [14] uses a particular folksonomy, RDF and ontologies to describe place names using tags. FODGS was developed to disambiguate and search place names from China. In FODGS, each place name has a unique footprint (geographic coordinate) and tag. However, this type of spatial representation has disadvantages, because various names with different meaning can have the same footprint. Another negative point in FODGS is its approach of computing spatial footprints offline. When a new footprint is updated, it is necessary to compile all spatial footprints in the FODGS data set and that operation takes a long time. This approach is not viable for use with VIG, because users will be updating the gazetteer all the time.
- The OntoGazetteer [15] uses ontology concepts to map data from web news sources to geographic data. Another related gazetteer [16] uses the same concepts as OntoGazetteer, but in the hydrograph name expansion context. Both gazetteers do not provide any structure for the use of VIG, semantics-based retrieval or LOD. Furthermore, both articles [15,16] do not state which ontologies were used to disambiguate place names.

[1] http://www.w3.org/TR/rdf-sparql-query/.

[2] http://splink.cria.org.br/.

Download English Version:

https://daneshyari.com/en/article/424550

Download Persian Version:

https://daneshyari.com/article/424550

Daneshyari.com