Journal of Biomedical Informatics 60 (2016) 177–186

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A method and software framework for enriching private biomedical sources with data from public online repositories

Alberto Anguita [a,*], Miguel García-Remesal [a], Norbert Graf [b], Victor Maojo [a]

[a] Biomedical Informatics Group, Universidad Politécnica de Madrid, Spain
[b] Department of Paediatric Oncology and Haematology, Saarland University Hospital, Germany

ARTICLE INFO

Article history:
Received 16 June 2015
Revised 1 February 2016
Accepted 3 February 2016
Available online 10 February 2016

Keywords:
Semantic integration
RDF
Public databases

ABSTRACT

Modern biomedical research relies on the semantic integration of heterogeneous data sources to find data correlations. Researchers access multiple datasets of disparate origin, and identify elements—e.g. genes, compounds, pathways—that lead to interesting correlations. Normally, they must refer to additional public databases in order to enrich the information about the identified entities—e.g. scientific literature, published clinical trial results, etc. While semantic integration techniques have traditionally focused on providing homogeneous access to private datasets—thus helping automate the first part of the research, and there exist different solutions for browsing public data, there is still a need for tools that facilitate merging public repositories with private datasets. This paper presents a framework that automatically locates public data of interest to the researcher and semantically integrates it with existing private datasets. The framework has been designed as an extension of traditional data integration systems, and has been validated with an existing data integration platform from a European research project by integrating a private biological dataset with data from the National Center for Biotechnology Information (NCBI).

1. Introduction

Semantic integration of disparate data sources has become in the last decade one of the pillars of biomedical research. Homogeneous access to heterogeneous data allows researchers to find novel correlations that find their application in a wide spectrum of areas such as improved diagnosis methods, more powerful and safer drugs, or personalized therapies [1–7]. To achieve this, extensive work has been devoted toward the development of architectures and systems capable of automatically integrating disparate sources [8]. Despite differences in approaches and architectures, all developed systems rely on the use of a schema that covers all integrated data, usually referred to as conceptual schema. Thus, each integrated data source is annotated with a description of semantic relationships that allow translating its contents to the conceptual schema. This information is referred to as database annotation. During the latest years, the RDF standard developed by the W3C (http://www.w3.org/RDF) has been widely used for defining the required conceptual schemas in data integration systems. This generalized adoption has allowed alleviating the problems derived from syntactic heterogeneities, and focusing on semantic divergences [9].

In most situations, data integration systems are employed to provide homogeneous access to a set of private databases, where the organizations owning those databases provide access to both the data and the schema of the datasets. Obviously, this approach limits the scope of the data integration system to the data that they obtain from such organizations. However, researchers often need to complement private datasets with data retrieved from public databases. Public databases are those that store free to access information. Most public databases provide access through the Internet with no restrictions, with web-based interfaces that can be accessed with an Internet browser. However, their schema is generally hidden to the outside world, and raw data is never exposed. As a consequence, data integration systems cannot automatically integrate public data records with instances from private repositories. Users who want to expand private repositories with public data are therefore forced to manually define queries in the respective public repositories—or a compatible public data search engine—for each concept they are working with. While this might be feasible if the user is working with few concepts, it becomes impractical once the private dataset involves more than a dozen concepts. The most prominent example of public repository in the biomedical domain is the data hosted by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)

* Corresponding author at: Group of Biomedical Informatics, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Spain. Tel.: +34 93 336 7467; fax: +34 91 352 4819.
E-mail address: aanguita@infomed.dia.fi.upm.es (A. Anguita).

http://dx.doi.org/10.1016/j.jbi.2016.02.004
1532-0464/

(NCBI). The NCBI has been providing relevant biomedical data to the scientific community for more than two decades now. The hosted data ranges from scientific publications—i.e. PubMed, PubMed Central—to genomic repositories—i.e. Gene, GEO Profiles.

Public databases generally provide a keyword-based search engine for users to issue queries. The user types a word, and the search engine returns all the items potentially related to that word. Consequently, when a user wishes to complement a private dataset with the information from a public dataset, she first identifies the terms of interest in the private dataset and then performs a query for each of them. This procedure is tedious and time-consuming, as the user is forced to issue multiple queries and navigate through different interfaces. Furthermore, and most important of all, the results of queries are volatile, in the sense that different users working with the same concepts will have to perform the same steps, and will not benefit from previous searches on similar concepts by their colleagues.

In this paper, we describe a framework for automatically enriching private datasets with related data from public repositories. This method has been conceived as an extension of heterogeneous data integration systems, and enables the semantic integration of private data with the related public data so that they can be homogeneously retrieved. As a consequence, our technique replicates in an automatic manner the procedure carried out by researchers to complement their working data with external information from public repositories. The researcher is therefore relieved from the hassle of manually searching the related public data and linking it to the private data.

Our work has allowed the semantic integration of NCBI data in a data integration platform provided by the p-medicine project. The latter is a European research project targeted at the development of an innovative technological platform for the management of post-genomic clinical trials [10]. It includes a layer that provides semantic integration of heterogeneous biomedical data sources. This layer adopts a centralized architecture governed by a data warehouse that stores the merged datasets in RDF triple format [11]. As conceptual schema, p-medicine uses the Health Data Ontology Trunk (HDOT), an ontology written in RDF covering the domain of cancer-related clinical trials [12]. Classes of the HDOT ontology are used to describe the entities stored in the data warehouse. In many situations, as with the micro RNAs classes mentioned later in the paper, RDF instances are used to refer to laboratory samples of a specific type—e.g. the RDF instance *hsa-let-7g-inst-1* that belongs to the class *hsa-let-7g* refers to a sample of this micro RNA from a single patient collected in a specific microarray experiment. At the same time, we use a wrapper that provides RDF based access to the NCBI repositories and that allows retrieving its data using SPARQL queries [13].

The paper is organized as follows: Section 2 expands on the related background. Section 3 describes our framework. Section 4 presents the evaluation performed on in a real scenario. Section 5 discusses on the benefits and originality provided by our contribution, and compares it to existing initiatives. Finally, Section 6 draws the conclusions and proposes future research lines.

2. Background

Although most efforts in the field of data integration have targeted private databases, there exist two noteworthy initiatives that focus on public repositories: Linked Data (http://linkeddata.org/) and Bio2RDF. The former is an initiative devoted to facilitating the access of public data in the Internet [14]. Linked Data promotes the publishing of data in RDF format, and provides guidelines and tutorials so that public datasets can be interlinked. Data owners are

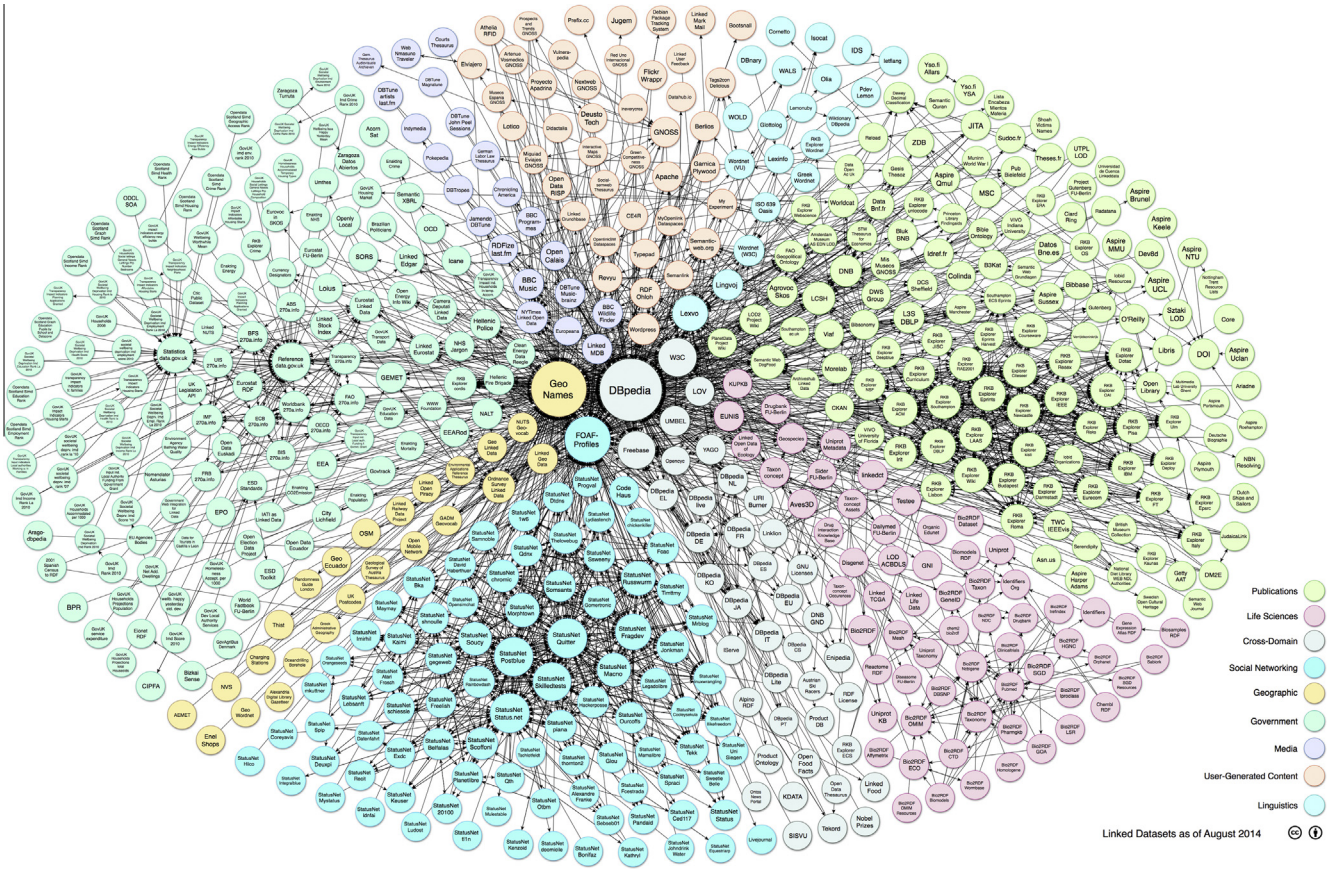

Fig. 1. Diagram representing the datasets that adopted the Linked Data publishing guidelines in 2014.