2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Curation of physical objects in botany:

Architecture and development of a linked open data-based application

Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, Adriana S. Vivacqua
Graduate Program in Informatics (PPGI)
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil
marcelayagui@ufrj.br, luisfmpm@ufrj.br, jonice@dcc.ufrj.br, avivacqua@dcc.ufrj.br

***Abstract*— Cultural heritage institutions store and manage large volumes of historical information, that have great material and humanitarian value. Their means of data organization, often obsolete, hinders the dissemination and reuse of information to be made in an effective way. In this sense, an application of Linked Open Data (LOD) technology is the possibility to extend the knowledge of a collection, with the use of open data already consolidated on the web to describe artworks or any type of physical object. This paper introduces a generic architecture based on an Extract, Transform, Load (ETL) methodology that connects LOD data from cultural heritage available on the web, generates descriptive content for physical objects and provides the mapping of research institutes engaged in studying them, highlighting their collaboration network. For instance, an implementation of the architecture was dedicated to the domain of Botany and can assist herbarium managers in creating exhibitions about medicinal plants. In this paper, we present a software architecture that provides an automatic method for creating dynamic pages from data stored in interconnected databases, and an application, which supports visitation systems with visualization and interaction mechanisms to encourage visitor learning.**

Keywords*— *Cultural Heritage; Linked Open Data; Georeferencing; Botany; RDFa.

I. INTRODUCTION

Cultural Heritage (CH) is a term that represents all the legacy of physical objects, environment, traditions and cultural knowledge of a society that are inherited from the past, maintained and developed in the present and preserved so that they can be enjoyed by future generations [1].

The web has become an increasingly important environment for publishing CH content of various types. For instance, many libraries provide digital content through their online repositories, museums exhibit their collections through collection browsers and documentation of all sorts of intangible heritage can be found in the form of different media types, such as audio and video recordings, interactive hypertext applications, or even, in a more playful way, in the form of games [1].

With the diffusion of web technologies and the digitalization of these resources, it has become important to develop new mechanisms to use and manage this large amount of information [2]. For this reason, many institutions, groups, and research projects that work in the domain of CH have developed online catalogs and repositories to manage and promote their heritage. In this context, the problem of semantic ambiguity between these catalogs and CH repositories is intensified due to the vast diversity of sampling occurrences and methods of obtaining the different types of documents and physical objects.

A solution to this problem is the creation of classification mechanisms and controlled vocabularies to define relationships between heterogeneous repositories and to add semantic value to semi-structured data scattered through the web [3]. These relationships, with semantic or conceptual denotation, can be used to reduce ambiguity. They help to better understand the meaning and build connections between online databases to bind open data to the public. These semantically enriched and connected data are also known as Linked Open Data (LOD) [4]. The semantic web and LOD technology have opened a new precedent for organizing and promoting heterogeneous content within different catalogs [5]. Supported by the Resource Description Framework (RDF), used to triplify these non-relational databases, and with the aid of the SPARQL query language, LOD technology makes it possible to assign semantic meaning to open data scattered on the internet, create databases supported in this content type, and to connect them with other existing databases such as DBpedia [6].

Cultural institutions can benefit from the adoption of open data access policies, increasing the use of collections and the number of potential users, because a problem faced by these organizations is the production of meaningful content for these artworks. Curators and museum managers are not encouraged to create interfaces that induce the learning and engagement of visitors [7]. The semantic web allied to LOD technology is an opportunity to organize and manage heterogeneous content within different catalogs of museums, libraries, research and cultural institutions. This way of managing data supports the diffusion of the CH in different contexts and creates more efficient and consistent means of communication [2]. Thus, the study of the semantic web and its mechanisms of association between things is necessary for the development of applications and intelligent information systems, enabling the integration of data with other systems to generate content in a more efficient and collaborative way.

This work is supported by LOD technology, making use of semantic web resources to build an application that aims to interconnect data from existing CH physical objects in

978-1-5386-1956-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.149

repositories available on the web so that relevant content can be delivered to visitors. The domain of museums, herbariums, and botanical institutes was chosen to describe information and provide visualization of the mapping of medicinal plants. Through the crossing of these data and a friendly interface, the application provides an efficient method for students, teachers and general users to access and visualize consolidated data about plants and institutions that have interest in disseminating content and CH.

II. Related Works

Applications that use geolocalization data supported by DBpedia and Geonames are based on the same approach that was used in this work. This is the case of the DBpedia Mobile application, which is focused on geolocalization supported by Dbpedia: the system retrieves recent georeferencing data from the mobile device and displays a map with information about nearby locations [6]. In Ariyani et al. [8], these databases were used to provide mechanisms capable of using geolocalization of an artwork to enrich users' experience in seeking relevant information about CH objects. Emaldi et al. [9] proposed an application based on QR Codes to provide additional data retrieved from LOD to detail artworks, in order to assist curators in the creation and maintenance of exhibitions. Our proposal connects consolidated databases in LOD to enhance the user experience when visiting an exhibition. Unlike the studies cited, which only use one interaction mechanism (e.g. interactive maps or QR Codes), the deployed application uses both approaches, as well as having a method that highlights the collaboration network between objects and institutes which research them.

Finally, the Global Biodiversity Information Facility (GBIF) provides a collaborative infrastructure for open data that allows unrestricted access to data on all kinds of life on the planet. GBIF works as an access portal that, assisted by its own development API, provides information on specimens, museum repositories, herbariums, research institutions and collaboration networks, as well as data of ocurrences and a visualization interface with species georeferencing [10]. Our application differs from GBIF in the way the data is presented. While GFIB uses a non-interactive interface for visualizing species maps, our application allows the user to interact with the map containing the species and their respective research institutes, in addition to displaying information enriched with GFIB's own metadata.

III. Architecture

The application's architecture is divided into four modules that, using the Extract, Transform, Load (ETL) tool Knime are responsible for (i) retrieving information from medicinal plants in LOD repositories and GBIF; (ii) transform the extracted data into (i); (iii) integrate these data; (iv) generate the web pages of each plant. At the end of the ETL process, the pages are published on a web server and a Quick Response Code (QR Code) corresponding to each plant is generated, allowing instant acccess from mobile devices. Fig. 1 illustrates the proposed architecture of this work and the components involved.

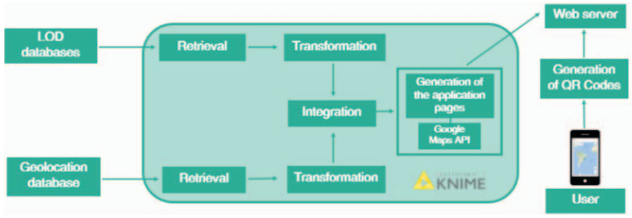

Fig. 1. Application Architecture

The architecture, the ETL process used, and the application implementation are explained below.

A. LOD databases

For this work, we used two LOD databases available on the web, DBpedia and Bio2RDF. DBpedia provides structured information from Wikipedia, enabling advanced queries to be performed on that data and also integrated with any other triplified databases [11]. DBPedia is used for data recovery of the popular name, synonyms, description, taxonomy, and image. Another LOD database used was Bio2RDF, a project that has the purpose of making available data related to life sciences in RDF format [12]. In this work, the Bio2RDF database was used to obtain data related to the Medical Subject Headings (MeSH) vocabulary, which integrates biomedical and pharmaceutical information, compatible with MEDLINE/PUBMed [13]. Data retrieval in these repositories was related to plant MeSH identifier and related annotations.

B. GBIF

Given that there were no LOD datasets related to plant geolocation (in particular, GeoSpecies), we used an open database for the retrieval of geolocalized plant data, the GBIF. From it, we collected geographical data related to plants and curating institutions that studied them, such as museums, herbariums, and research institutes. At the prototype level, these data were limited to only some occurrences in the Americas.

C. Architecture Modules

1) Retrieval

This module is responsible for retrieving data from the three databases mentioned. For DBpedia and Bio2RDF the data were retrieved with the implementation of SPARQL queries. This module is formed by DBpedia's endpoint connection activities, followed by a query that returns the following data: rdfs: label, dbo: abstract, dbo: thumbnail, dbo: kingdom, dbo: division, dbo: class, dbp: ordo, dbp: Family and dbp: synonyms, of the category "Medicinal_Plants" in English language. This query returned 945 plant records.

In addition, this module is composed by the connection to Bio2RDF's endpoint and the query that retrieves data referring to dcterms: title, mesh_vocabulary: mesh-scope-note and bio2rdf_vocabulary: identifier, of the category "Plants, Medicinal". The Bio2RDF database contains only English files, for this reason, it was not necessary to filter the database by language. This query returned 51 plant records.

On GBIF, data were extracted through a query in the virtual repository and converted to CSV format for later processing, to relate the occurrences of plants and the mapping of institutes, museums, and herbariums.

2) Transformation

In this module, all the information retrieved in DBpedia and Bio2RDF is transformed so that only data without RDF language format is used in the application. For the GBIF database, where data from plants and institutes are in CSV format, a filter was applied so that only correlational information between the plants retrieved at the LOD databases and corresponding institutes were used in the application map.

3) Integration

This module is responsible for integrating the previous data according to the standard identification of each database, i.e. the rdfs:label of DBpedia with dcterms:title of Bio2RDF and finally with GBIF's genus. This module generated, as output, data integration of 16 plants.

4) Generation of the application pages

Finally, this module is responsible for the application's front-end generation. For this purpose, after the transformation and integration of GBIF data, a JSON file containing the coordinates and other information that is shown on the map was generated.

With all the data available, generation of pages occurred by tagging of the triples retrieved in the language Resource Description Framework in Attributes (RDFa) (RDF code embedded in HTML), CSS Bootstrap Framework and JavaScript /Google Maps API (to build the map).

Knime was used to automate the 'Retrieval', 'Transformation', 'Integration', and 'Generation of the application pages' modules. Knime is an open source software based on ETL concepts that allows process creation and data transformation through integrated tools and advanced algorithms. It has more than 1,000 preloaded modules, discussion forums and hundreds of examples available for learning [14]. Knime was chosen because it allow the steps to be automated to generate all the pages through a code implementation in Java language, whichever is the number of pages and according to the amount of data recovered.

Thus, three model pages were implemented for the cascading generation: (i) the home page, (ii) the mapping of objects and institutes and (iii) specific data. To generate these three model pages to each of the objects, codes in the Java language that create these pages were implemented.

D. Publication of data on the Web

The generated pages were uploaded to the FTP server through loading in FileZilla, an open source software. Additionally, Google Analytics Javascripts were implemented on the application's pages to generate statistical data from it and to evaluate the users' electronic behavior and the application's performance.

After each document has been associated with a custom URI, it is possible to generate unique QR Codes for each of the application's plant record. This is necessary for usage in museums, herbariums or any other institution that needs a visitation system or some particular way of data visualization. Through the QR Code, visitors can use the camera of their mobile device to interact directly with the application, allowing the URI to open in a web browser and the page corresponding to the selected plant is displayed.

IV. USAGE SCENARIO

The application was tested on the operating system Android Lollipop 5.0.2 with a QR Code reader application. A typical application scenario would be a group of visitors at a plant exhibition, where each plant has its QR Code, that links to the page created with the data retrieved from LOD and GBIF. A visitor wanting more information about a plant, can use hisher smartphone equipped a QR Code reader and access the pages that contain the information related to the plant. For instance, Fig. 2 ilustrates the usage scenario for Lavandula.

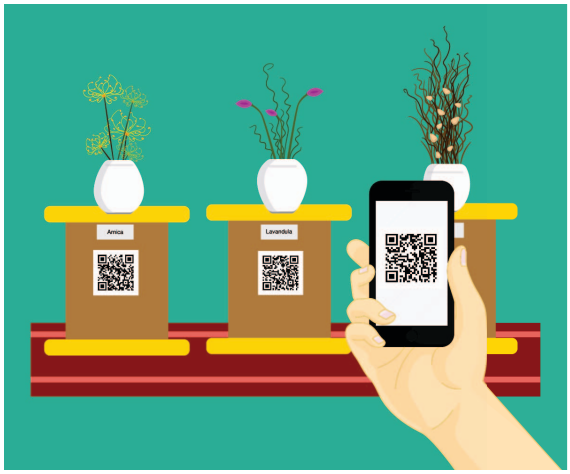

Fig. 2. Usage scenario

Reading the QR Code from Lavandula, the corresponding page with all the information retrieved from the LOD database will load, as shown in Fig. 3 (a). The information shown is the plant's image provided in DBpedia, its label (plant name), the annotation of the database Bio2RDF (subtitle 'About') and the abstract of the plant in DBpedia ('Description'). Additionally, navigation links are displayed for the other two pages, biological information ('Biology') and mapping of Species and institutes ('Mapping').

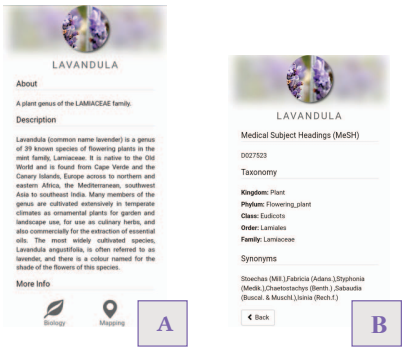

Fig. 3. (a) Home screen and (b) Biological information screen

In Fig. 3 (b) the biological information page is displayed, showing the MeSH identifier retrieved from Bio2RDF, the taxonomy and synonyms retrieved from DBpedia.

Finally, Fig. 4 (a) shows Lavandula's map. The points marked in green symbolize the species of plants mapped and the dots in red represent Institutes/Museums/Herbariums. The blue line linking two points represents the connection between the collected species and the institution responsible for its collection and curation. When touching a plant collection point (green), it is possible to open an informational balloon that displays plant's occurrence data such as name, location and catalog number, as well as an associated image, as shown in Fig. 4 (b). The Institutes/Museums/Herbariums points also have an informative balloon. These balloons show the type of institute, name, location, and number in the catalog, as shown in Fig. 4 (c).

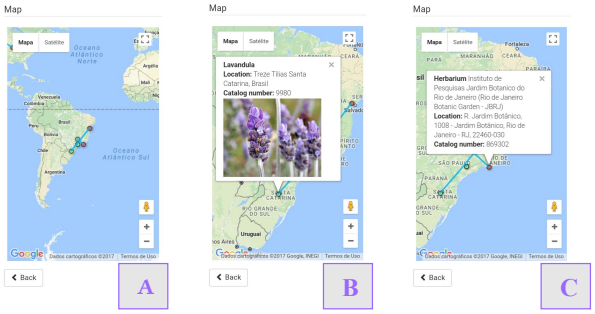

Fig. 4. (a) Mapping screen of plants and institutes, (b) Information balloon of a plant and (c) Information balloon of an institute

V. Conclusion

In this work we present all the steps involved in designing an architecture and implementation of an application that uses QR codes to make it possible to preview information and mapping plants, collecting locations and related research institutes, located in the American continent. For this purpose, the ETL tool Knime was used to implement the application by allowing the retrieval, transformation and integration of consolidated data extracted from the LOD and GBIF databases. In addition we automated the process of generating static and dynamic HTML pages, based on the integrated data of these databases, which also enables a large volume of pages through the proposed architecture to be created with minimal effort and time.

The application can be used as a support tool for museums, herbariums or any other institution that needs a visitation system or some particular way of visualizing the data. Through the QR Code, a visitor can use the camera of his mobile device to interact directly with the application, allowing the visualization of physical objects' information automatically. In addition, the architecture has been designed so that the processes implemented in Knime can be adapted to other types of cultural heritage and physical objects. The main contributions of our research are: (i) the creation of the architecture, with the automatic recovery, transformation and integration process that creates the application pages; (ii) the method to highlight the collaboration and curation network between the research institutes and their studied plants; (iii) the application itself, which supports visualization systems with visualization mechanisms and interaction to improve the learning and engagement of visitors.

In this way, it was possible to acknowledge some of the advantages that come from the usage of LOD technology and open acess data, including: (i) support in generating dynamic content and adding context to returned results; (ii) support in the creation of auxiliary tools for preservation and dissemination of CH; (iii) data integration with other systems, from triplified/published data on the semantic web; (iv) easier retrieval of information and public access to data already consolidated on the web.

For future work, it is intended to improve the application and carry out further studies to evaluate the user interface and engagement, among them: develop a method for automatic generation of QR Codes. Propose a method to evaluate the application, based on usability testing and on its web traffic, based on statistical data collected through Google Analytics, which was previously installed.

References

[1] E. Hyvönen, "Publishing and using cultural heritage linked data on the semantic web," *Synth. Lect. Semantic Web Theory Technol.*, vol. 2, no. 1, pp. 1–159, 2012.

[2] V. Ferrara, S. Sapia, A. Macchia, and F. Lella, "Cultural Heritage Open Data for Developing an Educational Platform," *Int J Comput Intell Stud*, vol. 5, no. 1, pp. 19–29, Apr. 2016.

[3] A. J. Warner, "A taxonomy primer," *Retrieved Sept.*, vol. 1, p. 2004, 2002.

[4] G. Smith, *Tagging: people-powered metadata for the social web*. New Riders, 2007.

[5] T. Berners-Lee, "Linked Data," *Design Issues*, 27-Jul-2006. .

[6] C. Becker and C. Bizer, "Exploring the Geospatial Semantic Web with DBpedia Mobile," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 7, no. 4, pp. 278–286, Dec. 2009.

[7] J. Marden, C. Li-Madeo, N. Whysel, and J. Edelstein, "Linked Open Data for Cultural Heritage: Evolution of an Information Technology," in *Proceedings of the 31st ACM International Conference on Design of Communication*, New York, NY, USA, 2013, pp. 107–112.

[8] N. F. Ariyani, M. B. Hanafi, H. Ginardi, and M. Saralita, "Linking spatial information of cultural heritage metadata with geo linked open data," in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2016, pp. 1–5.

[9] M. Emaldi, J. Lázaro, X. Laiseca, and D. López-de-Ipiña, "LinkedQR: Improving Tourism Experience through Linked Data and QR Codes," in *Ubiquitous Computing and Ambient Intelligence*, 2012, pp. 371–378.

[10] GBIF, "What is GBIF," *GBIF.ORG*, 19-Aug-2013. [Online]. Available: http://www.gbif.org/what-is-gbif. [Accessed: 06-Jan-2017].

[11] DBpedia, "About | DBpedia," 2017. [Online]. Available: http://wiki.dbpedia.org/about. [Accessed: 09-Jan-2017].

[12] M. Dumontier, "bio2rdf/bio2rdf-scripts," *GitHub*, 24-Jun-2016. [Online]. Available: https://github.com/bio2rdf/bio2rdf-scripts. [Accessed: 06-Jan-2017].

[13] Bio2RDF, "Bio2RDF::Mesh - the Datahub," 30-Jul-2016. [Online]. Available: https://datahub.io/dataset/bio2rdf-mesh. [Accessed: 06-Jan-2017].

[14] M. R. Berthold *et al.*, "KNIME-the Konstanz information miner: version 2.0 and beyond," *AcM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, 2009.