# Applying Linked Open Data and ETL for Mapping and Visualization of Physical Objects in Botany

Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, Adriana S. Vivacqua
Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI/UFRJ) – RJ – Brasil
{marcelayagui,luisfmpm}@ufrj.br, {jonice, avivacqua}@dcc.ufrj.br

## ABSTRACT

The purpose of this paper[1] is to show an architectural project to develop an application that aims to connect data from existing plants to available repositories on the Web, generate content and provide the mapping of these plants and museums, herbariums and research institutes engaged in studying them, highlighting their collaboration network. To this end, the ETL tool Knime was used, with the aid of LOD technology and open data extracted from GBIF, in order to provide an automatic method for creating dynamic pages where information of medicinal plants can be viewed and the mapping of the collection linked to the plant image in association with the related research institutes, providing users with a wide view of the georeferenced area.

## KEYWORDS

Cultural Heritage; Linked Open Data; Georeferencing; RDFa.

## 1 INTRODUÇÃO

Museus, galerias, herbários e diversas instituições do Patrimônio Cultural (PC) armazenam e gerenciam grande volume de informações históricas, que possuem grande valor material e humanitário. Seus meios de organização dos dados, muitas vezes obsoletos, impedem que a divulgação e reutilização das informações seja feita de maneira efetiva [6]. Contudo, com a evolução da web dos documentos, e recentemente, da web semântica, algumas dessas instituições perceberam a importância de disponibilizar seus repositórios de dados no formato aberto, além de conectá-los com outras bases compatíveis com padrões de metadados e ontologias existentes, adicionando valor semântico aos dados semiestruturados que até então, eram armazenados localmente. Com este tipo de organização das informações, conhecido como *Linked Open Data* (LOD), é possível conectar as diversas fontes de dados heterogêneas tornando a web num enorme repositório de dados global [3].

Por meio da tecnologia LOD, é possível agregar valor à recursos da web com a realização de consultas que navegam nas conexões formadas entre essas fontes de dados. No domínio do PC, uma aplicação de LOD é a possibilidade de ampliar o conhecimento de uma coleção, com a utilização dos dados abertos já consolidados na web para descrever obras de arte ou qualquer tipo de objeto físico. Com a recuperação de informações disponíveis nessas bases e através do agrupamento de conceitos que se conectam entre si, é possível gerar conteúdo relevante para usuários que visitam essas galerias.

Esta abordagem representa uma oportunidade de auxiliar curadores e gestores de museus na criação e manutenção de exposições, pois um problema enfrentado nessas instituições é a produção de conteúdo descritivo para obras. As interfaces criadas por estes profissionais não estimulam o aprendizado e engajamento de visitantes. A tecnologia LOD é uma oportunidade para organizar e gerenciar conteúdo heterogêneo dentro de diferentes catálogos de instituições culturais. Esta forma de gerir os dados apoia a difusão do PC em diferentes contextos e cria meios de comunicação mais consistentes [1, 5].

Este trabalho se apoia na tecnologia LOD para construir um aplicativo que tem como finalidade interligar dados de objetos físicos do PC existentes em repositórios disponíveis na web para que conteúdo relevante seja gerado a visitantes. Neste caso de aplicação, foi escolhido o domínio de museus, herbários e institutos relacionados à botânica para descrever informações e fornecer visualização do mapeamento de plantas medicinais. Através do cruzamento destes dados, o aplicativo oferece um método eficiente para que usuários em geral possam acessar e visualizar dados consolidados acerca plantas e instituições que possuem interesse em divulgar conteúdo e PC.

## 2 ARQUITETURA

A arquitetura do aplicativo é dividida em quatro módulos que, com o auxílio da metodologia *Extraction, Transformation, Loading* (ETL) [4] e da ferramenta Knime são responsáveis por (i) recuperar informações de plantas medicinais em repositórios LOD (DBpedia e a Bio2RDF) e da base de dados abertos no padrão três estrelas de Tim Berners-Lee[2] para a recuperação de dados geolocalizados de plantas (*Global Biodiversity Information Facility* - GBIF); (ii) transformar e limpar os dados extraídos em (i); (iii) integrar esses dados, a partir do rdfs:label da DBpedia com o dcterms:title da Bio2RDF e com o '*genus*' do GBIF. Neste estudo, foi resultante deste módulo a integração dos dados de 16 plantas; (iv) gerar as páginas web de cada planta, na linguagem *Resource Description Framework in Attributes* (RDFa) (código RDF embutido no HTML), com o *framework* CSS Bootstrap e com

[2] https://www.w3.org/DesignIssues/LinkedData.html

In: Sessão de Pôsteres do WebMedia´17, Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN: 978-85-7669-380-2.

JavaScript/API *Google Maps* para a criação dos Mapas. Três páginas-modelo foram implementadas na linguagem Java para a geração em cascata, sendo a primeira a página inicial, a segunda com o mapeamento das espécies e institutos e a terceira com dados biológicos mais específicos. Ao final do processo ETL as páginas foram publicadas em um servidor web e um QR *Code* correspondente a cada planta foi gerado, possibilitando a identificação instantânea a partir de dispositivos móveis. A Figura 1 ilustra a arquitetura.

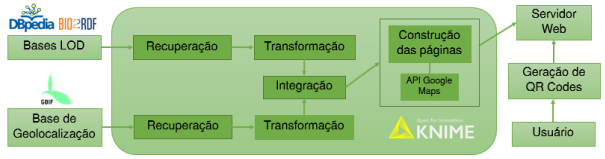


**Figura 1: Arquitetura do aplicativo.**

## 3 APLICATIVO

Para que um usuário possa visualizar informações sobre uma planta e/ou seus locais de coleta e instituições curadoras em dispositivo móvel, é necessário ter um aplicativo instalado que suporte a leitura de QR *Codes* para escanear o código correspondente àquela planta e abrir o link disponibilizado. A nível de protótipo e para fins de demonstração, foi criada uma página[3] que contém todos os QR Codes com links para as respectivas plantas.

A exemplo disso, ao ler um QR *Code* será carregada a página com dados recuperados das bases LOD. As informações mostradas são a imagem da planta disponibilizada na DBpedia, seu *label*, a anotação da base Bio2RDF ('*About*') e o *abstract* da planta na DBpedia ('*Description*'). Além disso, são exibidos links de navegação para outras páginas, como informações biológicas ('*Biology*') e mapeamento de espécies e institutos ('*Mapping*'). A segunda página consiste em mostrar dados biológicos, onde são exibidos o identificador do *Medical Subject Headings* recuperado do Bio2RDF, a taxonomia e sinônimos recuperados da DBpedia.

Por fim, há a página de mapeamento de espécies e institutos, onde os pontos marcados em verde simbolizam as espécies de plantas mapeadas e os pontos em vermelho representam Institutos/Museus/Herbários. A linha azul que liga dois pontos representa a conexão entre a espécie coletada e a instituição responsável por sua coleta e curadoria. Ao tocar um ponto de coleta de planta (verde), é possível abrir um balão informativo que exibe dados sobre a ocorrência da planta, como nome, localização e número no catálogo (GBIF), bem como uma imagem associada (DBpedia). Os pontos de Institutos/Museus/Herbários também possuem um balão informativo. Estes balões mostram o tipo de instituto, o nome, a localização e o número no catálogo.

## 4 TRABALHOS RELACIONADOS

Aplicativos que utilizam dados de geolocalização apoiados na DBPedia e no Geonames foram baseados na mesma abordagem que foi empregada neste trabalho, como é o caso do aplicativo DBPedia *Mobile*, onde o sistema recupera dados de georreferenciamento recentes do dispositivo móvel e exibe um mapa com informações recuperadas da DBPedia sobre locais próximos [1]. O aplicativo de [2] é baseado em QR *Codes* e fornece aos usuários dados adicionais que detalham uma obra do PC. Por fim, o GBIF é uma iniciativa que fornece uma infraestrutura colaborativa de dados abertos e oferece informações sobre espécimes, repositórios de museus, herbários e instituições de pesquisa, além de uma interface de visualização não interativa com georreferenciamento das espécies [7]. Diferentemente dos estudos relacionados citados, que só utilizam um mecanismo de interação (por exemplo, por meio de mapas ou QR *Codes*), o aplicativo implementado neste trabalho mescla ambas as abordagens, além de possuir um método que destaca a rede de colaboração entre objetos e institutos que os pesquisam. O aplicativo permite também que o usuário interaja com o mapa contendo as espécies e seus respectivos institutos de pesquisa, além de exibir informações enriquecidas com metadados do próprio GFIB.

## 5 CONCLUSÃO

Neste trabalho apresentamos a arquitetura de um aplicativo que utiliza QR *Codes* para tornar possível a visualização de informações e o mapeamento de plantas, locais de coleta e institutos de pesquisa relacionados. A ferramenta ETL Knime foi utilizada para implementação do aplicativo por permitir a recuperação, transformação e integração de dados consolidados extraídos das bases LOD e GBIF, além de automatizar o processo de geração de páginas HTML estáticas e dinâmicas, baseado nos dados integrados dessas bases, o que também possibilita que, através da arquitetura proposta, um grande volume de páginas seja criado com esforço e tempo mínimos. Além disso, a arquitetura foi projetada de modo que os processos implementados no Knime possam ser adaptados para outros tipos de PC e objetos físicos. Para trabalhos futuros é pretendido realizar incrementos no aplicativo e realizar estudos posteriores para avaliar a interface e o engajamento dos usuários, além de desenvolver um método para geração automática de QR *Codes*.

[3] https://goo.gl/moZDn2

## REFERENCES

[1] Becker, C. and Bizer, C. 2009. Exploring the Geospatial Semantic Web with DBpedia Mobile. *Web Semantics: Science, Services and Agents on the World Wide Web*. 7, 4 (Dec. 2009), 278–286.

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

[3] Heath, T. and Bizer, C. 2011. *Linked Data:Evolving the Web into a Global Data Space*. Morgan & Claypool.

[4] Kimball, R. and Caserta, J. 2011. *The Data Warehouse?ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. John Wiley & Sons.

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

[6] Ruthven, I. and Chowdhury, G.G. 2015. *Cultural Heritage Information: Access and management*. Facet Publishing.

[7] What is GBIF: 2013. *http://www.gbif.org/what-is-gbif*. Accessed: 2017-01-06.