Arquitetura para cocuradoria de dados de conhecimento popular integrados por meio de *Linked Open Data*

Marcela Mayumi Mauricio Yagui, Adriana S. Vivacqua

Universidade Federal do Rio de Janeiro – Rio de Janeiro, RJ – Brasil

marcelayagui@ufrj.br, avivacqua@dcc.ufrj.br

Abstract. *Popular knowledge databases are becoming essential for the preservation of a region's culture. In the Brazilian scenario, many museums and galleries don't incorporate systems that exploit collective intelligence in order to aid the recording of empirical information. The goal of this work is to present an architecture to support the co-curation of data derived from popular knowledge. Our proposal also provides the interconnection between visitors' contributions with data already available in open repositories on the web (in the Linked Open Data format). From this, it is hoped to achieve the enrichment of the cultural heritage of museums and their galleries, benefiting the community with access to enriched and cured knowledge databases.*

Resumo. *Bases de conhecimento popular estão se tornando essenciais para a preservação da cultura de uma região. No cenário Brasileiro, muitos museus e galerias não incorporam sistemas que exploram a inteligência coletiva de modo a auxiliar o registro de informações empíricas. Este trabalho tem como objetivo apresentar uma arquitetura para apoiar a cocuradoria de dados advindos do conhecimento popular. Nossa proposta também propicia a interconexão entre contribuições de visitantes com dados já dispostos em repositórios abertos na web (no formato Linked Open Data). A partir disso, espera-se conseguir o enriquecimento do patrimônio cultural de museus e suas galerias, beneficiando a comunidade com acesso a bases de conhecimento enriquecido e curado.*

1. Introdução

O *crowdsourcing* é um campo da *Crowd Computing* que surgiu como solução para problemas de gestão da informação por meio de geração de conteúdo por usuários na web social, onde há mobilizações e contribuições de multidões virtuais para encontrar e reunir informações. Através do *crowdsourcing* pode ser adotada a abordagem *co-curation* (em português, cocuradoria), específica para contextos aplicados ao Patrimônio Cultural (PC). A cocuradoria é derivada do termo curadoria, mas aplicada em ambientes colaborativos on-line, onde os participantes dessas atividades de cura fazem parte de uma multidão virtual [Oomen and Aroyo 2011].

Por meio do *crowdsourcing* – e consequentemente, da cocuradoria – é possível fomentar a participação do público para construir bases de conhecimento popular, que atualmente são indispensáveis para a preservação do PC de uma região. Para Cotterill *el al.* (2016) o registro do conhecimento e a digitalização de informações podem prolongar

a vida útil de objetos e histórias, tornando-as conhecidas por outras pessoas, como também preservadas para as gerações futuras.

A fim de apoiar a criação e manutenção de bases de conhecimento popular, este trabalho propõe uma arquitetura que apoia as atividades de cocuradoria deste tipo de PC, por meio da criação de chamadas *crowdsourcing* direcionadas à geração de conteúdo por visitantes. Além disso, esses dados são interligados com outros já dispostos em repositórios *Linked Open Data* (LOD), ampliando as possibilidades de recuperação, integração e publicação de conteúdo on-line. Neste trabalho buscou-se contribuir com a definição de uma arquitetura que reúne fontes de dados heterogêneas (conjunto de dados LOD e dados de contribuições de usuários) com a respectiva curadoria dessas informações.

2. Trabalhos relacionados

Estudos que utilizam a cocuradoria para assegurar o registro de informações de objetos do PC baseiam-se na mesma abordagem empregada neste trabalho. Como é o caso da plataforma Co-Curate, que integra contribuições de usuários com diversos tipos de materiais *open access* em um ambiente de museu virtual [Cotterill et al. 2016]. O trabalho de Rotman *et al.* (2012) apresentou um estudo que teve como objetivo entender as principais características de sistemas *crowdsourcing* relacionados à curadoria de conteúdo, destacando em sua análise o caso da Encyclopedia of Life. Diferente dos estudos mencionados, nosso trabalho utiliza técnicas da web semântica para interligar bases de dados e permitir a interoperabilidade entre sistemas, de modo a ampliar a recuperação de dados curados, além de permitir que o modelo de dados seja extensível.

Com relação às plataformas que utilizam *crowdsourcing* e LOD, pode-se citar o projeto Linked Jazz, que aplica a tecnologia LOD para curadoria de materiais de arquivamento digital no domínio das artes musicais. No projeto foram utilizadas técnicas de processamento de linguagem natural para reconhecimento e extração da rede de artistas da DBpedia, e o aprimoramento das conexões entre os músicos por meio de *crowdsourcing* [Pattuelli and Miller 2015]. O estudo de Pattuelli e Miller (2015) cura apenas material disponível em LOD e permite o aprimoramento desse material, não suportando a criação ou a adição de novos conteúdos. Nossa proposta utiliza o *crowdsourcing* com a abordagem *co-curation* em conjunto com a tecnologia LOD, para criar e curar diversos tipos de materiais com infraestrutura aberta.

3. Arquitetura

Esta seção apresenta a arquitetura proposta para este trabalho, que é dividida em cinco camadas: usuário, aplicação, transformação, integração e dados, explicadas a seguir. A Figura 1 ilustra a arquitetura e identifica os principais componentes envolvidos.

Camada de usuário: É composta pelos componentes ‘Contribuintes’ e ‘Curadores’. O componente ‘Contribuintes’ representa os usuários que visitam um museu e que têm a possibilidade de utilizar um dispositivo móvel para escanear um *QR Code* e recuperar informações sobre um objeto de uma ‘Coleção curada’. O componente ‘Curadores’ é formado pelos usuários especialistas que organizam de forma colaborativa assimétrica o conteúdo gerado pelos visitantes. Um usuário ‘Curador’ cria uma chamada *crowdsourcing* contendo o tema de uma coleção que fica aberta para ‘Contribuições’.

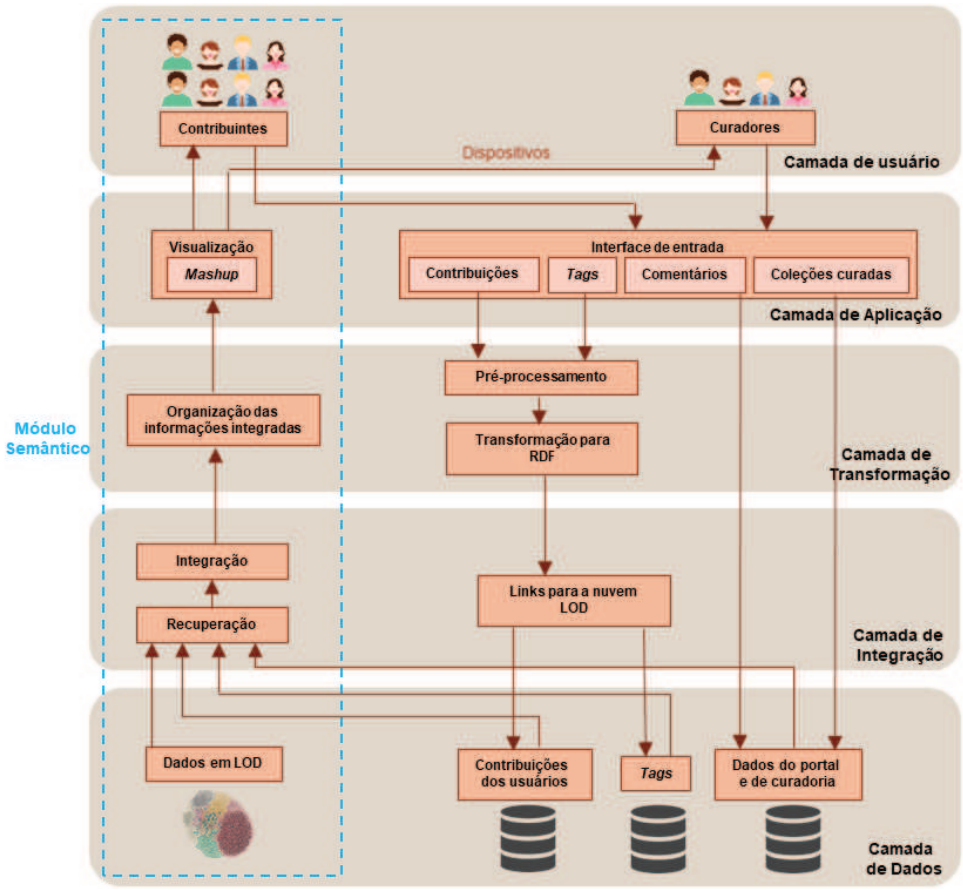

Figura 1 - Arquitetura do sistema

Camada de aplicação: É responsável pela geração do *front-end* da aplicação. Esta camada é dividida nos componentes 'Interface de entrada' e 'Visualização'. O componente 'Interface de entrada' é responsável por fornecer um meio pelo qual os usuários interagem com o sistema para a entrada de dados. Os usuários 'Contribuintes' fazem a entrada de dados das 'Contribuições', dos 'Comentários' e das '*Tags'* (componentes da arquitetura). As 'Contribuições' podem ser expressas em texto ou imagem, podendo ter tipos variados, como por exemplo, impressões pessoais, histórias, curiosidades sobre o tema da coleção ou qualquer outro tipo de material. Ou seja, por meio de *microtasking* (envio de pequenas tarefas/contribuições) o conteúdo é gerado pelo usuário com o envio de objetos para determinadas coleções.

Os 'Curadores' fazem a organização das informações com base em sua experiência sobre o tema da coleção, como indicado no componente 'Coleções curadas'. Eles analisam cada pequena 'Contribuição' para evitar que informações incorretas sejam adicionadas às coleções. Se um objeto estiver válido, ele fica disponível entre os objetos candidatos que podem ser selecionados no momento da curadoria. Caso contrário, o autor do objeto é notificado para realizar a correção de sua 'Contribuição'.

Após a validação das 'Contribuições', inicia-se o processo de curadoria das coleções. A primeira fase do processo está relacionada com a busca de conteúdo relevante para um determinado tópico escolhido para curadoria, sendo que este conteúdo é recomendado preferencialmente a partir das 'Contribuições' dos visitantes e dos dados dispostos em bases LOD. Para facilitar a busca dessas informações, são utilizadas as '*Tags*'

que foram associadas às 'Contribuições'. A partir disso, o 'Curador' seleciona, define e reorganiza as 'Coleções' usando sua experiência, ou seja, adiciona conexões entre os objetos para atribuir sentido à coleção.

Outro componente da camada de aplicação é chamado 'Visualização' (presente no Módulo Semântico). Neste componente ocorre a disponibilização das interfaces utilizando o '*Mashup'* de dados selecionados pelos 'Curadores', onde essas informações são estruturadas em *Resource Description Framework in Attributes* (RDFa) (código RDF embutido no HTML). Por meio de *QR Codes*, é possível acessar cada página e visualizar as 'Coleções'. Por meio deste componente, um 'Contribuinte' poderá reiniciar o processo de criação de conteúdo, e os 'Curadores' podem realizar novamente as tarefas de cura das 'Coleções'.

Camada de transformação: Esta camada tem como objetivo preparar os dados para posterior carga no banco de dados (componentes 'Pré-processamento' e 'Transformação para RDF') e na organização de interfaces obtidas por meio da integração dos dados (componente 'Organização das informações integradas').

O 'Pré-processamento dos dados' é o componente no qual os dados das 'Contribuições' e das '*Tags*' são analisados e limpos, sendo eliminados caracteres especiais, de modo a assegurar a qualidade das informações. Após o 'Pré-processamento', os dados são transportados para o componente 'Transformação' e convertidos em triplas RDF, com a sintaxe sujeito-predicado-objeto, onde o sujeito é uma URI, o objeto pode ser uma URI, um *blank node* ou um literal e o predicado é uma URI que define o relacionamento entre sujeito e predicado. O RDF foi utilizado pois favorece a integração de dados de diferentes esquemas, além de ser extensível, por suportar a evolução do modelo sem necessitar a alteração dos dados ou da sua estrutura.

O componente 'Organização das informações integradas' (presente no Módulo Semântico) consiste na geração de uma interface de visualização das informações curadas (provenientes das bases de dados integradas), que posteriormente são exibidas na camada de aplicação.

Camada de integração: É composta pelos componentes 'Links para a nuvem LOD', 'Recuperação' e 'Integração'. No componente 'Links para a nuvem LOD', são criadas ligações entre as triplas RDF com elementos da nuvem LOD por meio do Apache Stanbol. O Stanbol é um software *open source* cuja principal finalidade é acrescentar serviços semânticos em sistemas de gerenciamento de conteúdo, fornecendo componentes que processam conteúdo de linguagem natural em metadados RDF. Após as ligações criadas, os dados são enviados aos seus respectivos repositórios locais.

O componente 'Recuperação' (presente no Módulo Semântico) é responsável pela recuperação de dados das quatro bases que estão dispostas na camada de dados. Para a base LOD, base de 'Contribuição' dos usuários e base de '*Tags*' os dados são recuperados com a implementação de consultas federadas na linguagem SPARQL. Para a base de 'Dados do portal e de curadoria', os dados são extraídos por meio de consulta SQL no repositório relacional para posterior processamento e integração.

O componente 'Integração' (presente no Módulo Semântico) tem como objetivo integrar os 'Dados do portal e de curadoria', dispostos no formato relacional, com os dados em RDF e em LOD obtidos por meio de consultas SPARQL.

Camada de dados: É composta por três bases de dados armazenadas localmente e uma base de dados externa. A primeira base contém 'Dados do portal, administrativas do sistema e de curadoria' e 'Comentários', e é armazenada no formato relacional no SGBD MySQL. A segunda base corresponde aos dados das 'Contribuições' dos usuários, dos objetos textuais e imagens. A terceira base contém as '*Tags*' que descrevem as 'Contribuições', a fim de enriquecer as descrições e melhorar a recuperação dos dados. A segunda e terceira base são armazenadas em RDF no banco de triplas Apache Jena Fuseki. Por fim, a base externa é composta pela nuvem LOD, disponível na web, onde informações relacionadas aos objetos de contribuições são selecionadas posteriormente.

O Módulo Semântico [Yagui et al. 2017b] (destacado em azul) está indicado na arquitetura por representar o caminho pelo qual os dados em LOD percorrem até a formação de um conteúdo curado estar disponível para ser consumido por um usuário.

4. Prova de conceito

Para a aplicação prática do Módulo Semântico, foram utilizadas bases de dados abertas, disponíveis na web, relacionadas a plantas medicinais e instituições curadoras relacionadas à Botânica. Nosso objetivo foi testar o Módulo Semântico para realizar a curadoria de conteúdo aberto e disponibilizá-la por meio de um aplicativo web, de modo a permitir que visitantes possam consumir as informações curadas.

O aplicativo recupera e integra dados de variadas fontes, entre elas: a DBpedia, o Bio2RDF e o Global Biodiversity Information Facility (GBIF). O GBIF possui um acervo similar à proposta apresentada deste trabalho, de modo que ao realizar testes no Módulo Semântico com seus dados, esperou-se que o resultado pudesse ser o mais fidedigno possível a uma aplicação real.

O aplicativo foi testado de modo presencial em duas ocasiões diferentes, entre os meses de outubro [Yagui et al. 2017a] e novembro [Yagui et al. 2017b] de 2017. Em ambas as ocasiões, o aplicativo foi apresentado a visitantes em exposições do tipo 'feira de ciências'. Para avaliar a experiência dos visitantes, nós os convidamos a expressar sua opinião durante e após a utilização. Em geral, o aplicativo foi bem aceito: em particular, os usuários apreciaram a facilidade em acessar informações das plantas a partir do

País	Aquisição: Usuários
	83 (Porcentagem do total: 100,00% (83))
1. Brazil	48 (57,83%)
2. United States	13 (15,66%)
3. (not set)	5 (6,02%)
4. Italy	3 (3,61%)
5. Argentina	2 (2,41%)
6. Algeria	2 (2,41%)
7. Russia	2 (2,41%)
8. Côte d'Ivoire	1 (1,20%)
9. Colombia	1 (1,20%)
10. United Kingdom	1 (1,20%)
11. Greece	1 (1,20%)
12. India	1 (1,20%)
13. Mozambique	1 (1,20%)
14. Romania	1 (1,20%)
15. South Africa	1 (1,20%)

Figura 2 – Frequência de utilização do aplicativo

leitor de *QR Code*. Em contraste, foram relatadas algumas dificuldades para acessar informações de espécimes e institutos curadores no mapa. Adicionalmente, ao final das exibições disponibilizamos uma página que contém todos os *QR Codes* com links para as plantas e convidamos os usuários a compartilhar o aplicativo em suas redes sociais. Deste modo, foi possível obter dados sobre a utilização do aplicativo de modo não presencial, através de JavaScripts do Google Analytics adicionalmente implementados. Embora nesta última modalidade não tenham sido realizadas experimentações acerca da satisfação dos usuários, foi possível constatar que houve interesse pelo *App*, tendo o mesmo sido utilizado por diversos usuários, conforme mostra a Figura 2.

Detalhes adicionais sobre o Módulo Semântico estão relatados em [Yagui et al. 2017b].

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma arquitetura para dar suporte à criação e à manutenção de bases de conhecimento popular. Com a finalidade de apoiar as atividades de cocuradoria deste tipo de PC, a arquitetura faz uso de dois mecanismos de interação: (i) criação de chamadas *crowdsourcing* direcionadas à criação de conteúdo por visitantes e (ii) cura desses materiais por especialistas. A arquitetura também permite que os dados sejam interligados com outros já dispostos em repositórios LOD, ampliando a possibilidade de recuperação de conteúdo relacionado.

Além disso, foi realizada uma prova de conceito do Módulo Semântico da arquitetura com aplicação de bases de dados abertas criadas colaborativamente no domínio da Botânica. Desta forma, a proposta deste módulo mostrou-se adequada para o cenário de aplicação no qual este trabalho foi projetado, de modo que há suporte para a inclusão futura de outros dados empíricos gerados por visitantes. Como trabalhos futuros, serão realizados estudos baseados em testes de usabilidade e avaliações qualitativas com curadores de instituições culturais.

Referências

Cotterill, S., Hudson, M., Lloyd, K., et al. (2016). Co-curate: Working with Schools and Communities to Add Value to Open Collections. *Journal of Interactive Media in Education*, v. 2016, n. 1.

Oomen, J. and Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In *Proceedings of the 5th International Conference on Communities and Technologies*.

Pattuelli, M. C. and Miller, M. (2015). Semantic network edges: a human-machine approach to represent typed relations in social networks. *Journal of Knowledge Management*, v. 19, n. 1, p. 71–81.

Rotman, D., Procita, K., Hansen, D., Parr, C. S. and Preece, J. (2012). Supporting content curation communities: The case of the Encyclopedia of Life. *Journal of the American Society for Information Science and Technology*, v. 63, n. 6, p. 1092–1107.

Yagui, M. M. M., Maia, L. F. M. P., Oliveira, J. and Vivacqua, A. S. (2017a). Applying Linked Open Data and ETL for Mapping and Visualization of Physical Objects in Botany. In *Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres*.

Yagui, M. M. M., Maia, L. F. M. P., Oliveira, J. and Vivacqua, A. S. (2017b). Curation of Physical Objects in Botany: Architecture and Development of a Linked Open Data-Based Application. In *Proceedings of the 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(DASC/PiCom/DataCom/CyberSciTech)*. IEEE.