Linked Open Archives (LOA): Uma experiência de Interação em Grafos para Acervos Arquivísticos na Web

Rodrigo Oliveira[1] , Jair Martins de Miranda[2] , Alison Filgueiras[3]

[1]Programa de Pós-Graduação em Computação (PPGC) da Universidade Federal Fluminense (UFF), Niterói, RJ, Brasil

[2]Departamento de Arquivologia do Centro de Ciências Humanas e Sociais da Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ, Brasil

[3]Universidade Estadual de Goiás (UEG), Rio de Janeiro, RJ, Brasil

rodrigoso@id.uff.br, jairmm@unirio.br, alison.filgueiras@ueg.br

Abstract. *This paper presents the prototype of the Linked Open Archives (LOA) platform, which combines knowledge graphs and semantic web technologies to organize and visualize archival collections based on the Records in Contexts (RiC-O) ontology. LOA enables the interactive exploration of historical documents, connecting entities and relationships, and enabling search and navigation with interactive graphs and SPARQL queries. The research provides insights into using interactive graphs in human-data visualization and interaction.*

Resumo. *O artigo apresenta o protótipo da plataforma Linked Open Archives (LOA), que combina grafos de conhecimento e tecnologias da web semântica para organizar e visualizar acervos arquivísticos a partir da ontologia Records in Contexts (RiC-O). O LOA permite a exploração interativa de documentos históricos, conectando entidades e relações e permitindo a busca e navegação com grafos interativos e consultas SPARQL. A pesquisa oferece insights sobre o uso de grafos interativos na visualização e interação humano-dados.*

1. Introdução

Grafos de conhecimento são definidos como gráficos de dados que acumulam e transmitem conhecimento do mundo real. Os nós em gráficos de conhecimento representam as entidades de interesse, e as arestas representam as relações entre as entidades (Peng et al., 2023). Seu principal objetivo é descrever objetos de interesse e conexões entre eles para uso em diversas aplicações (Noy et al., 2019) (Maciel et al., 2024). Na busca ou recuperação de informações eles podem ser úteis para auxiliar o usuário a coletar informações, ou então percorrer o grafo para compreender as relações exibidas, inferir novos fatos e construir assim conhecimento, diferentemente de uma busca tradicional. Dessa forma, os grafos interativos são importantes ferramentas para avançar o estado da arte em interação de dados humanos e visualização de dados, já que a maioria da literatura relevante se concentra em dados tabulares (Sabou et al., 2016).

O uso de grafos é uma alternativa ao processo de recuperação da informação com uso de tecnologias convencionais de busca e visualização de resultados. Muitas

destas tecnologias inovadoras se baseiam no conceito da web semântica e dos dados abertos conectados (*Linked Open Data* – LOD). Berners-Lee et al. (2001) foram pioneiros nas tecnologias básicas para uma web semântica, sejam em protocolos e frameworks (RDF e RDF-S), ontologias e a linguagem OWL e a linguagem de recuperação baseada em SQL denominada SPARQL, como uma especificação para desenvolver sistemas de organização do conhecimento para a nova web (Berners-Lee, Hendler, Lassila, 2001). O objetivo dessa web semântica é a manutenção de informações com conteúdo semântico que possam ser processadas por máquinas e interligadas por computadores. Isto é, por meio de uma linguagem de marcação, a exemplo do RDF, que opera mediante triplas, compostas dos elementos "Sujeito, Propriedade, Objeto" e por serem semântica e univocamente identificadas na web através de uma URI (*Uniform Resource Identifier*)[1], que são as bases para gerar os grafos que interligam esses elementos.

Nesta pesquisa exploratória, experimental e aplicada apresentamos a plataforma *Linked Open Archives* (LOA) (www.linkedopenarchives.com), uma arquitetura que abriga tecnologias disruptivas na arquivologia se utilizando de técnicas de paleografia digital, ontologias, web Semântica e LOD com potencial para causar grande impacto nas práticas arquivísticas e no acesso aos arquivos. O LOA tem como meta inicial a disponibilização do conjunto documental Série Escravidão do Arquivo Geral da Cidade do Rio de Janeiro (AGCRJ) na web, via a ontologia RiC (ICA/EGAD, 2023). Visando futuramente a conexão com outros conjuntos documentais de Angola e Portugal sobre o Tráfico Atlântico de Escravos (*Atlântic Slave Trade*) no ambiente *Linked Open Data*.

Este artigo apresenta um relato da experiência do LOA quanto uma ferramenta para interação humano-dados na arquivologia atual por meio de grafos interativos e sua estrutura baseada na web semântica. A Seção 2 apresenta os conceitos e trabalhos anteriores que fundamentam esta pesquisa. A Seção 3, destaca a arquitetura e metodologia envolvida na plataforma. Na Seção 4 é apresentado a estrutura, modelo de interação e exemplos dos grafos interativos disponibilizados no LOA. Em seguida, a Seção 5, apresenta as conclusões do estudo e perspectivas de trabalhos futuros.

2. Fundamentação Teórica

2.1. Web Semântica

A Web Semântica é um arcabouço de tecnologias voltadas para uma extensão da web tradicional. Ela se baseia na criação de vocabulários controlados com o intuito de organizar e armazenar a informação, de forma que sua recuperação seja realizada de acordo com esse vocabulário definido, e assim, facilitar o trânsito de informações semânticas na rede (Berners-Lee, Hendler, Lassila, 2001). O principal objetivo é, através da interligação entre significados de palavras, conseguir atribuir um sentido aos conteúdos publicados na internet, extrapolando a interpretação dos dados apenas pelo olhar humano, possibilitando que máquinas possam processá-los. O ponto de partida é o domínio, dentro dessa perspectiva, da OWL (*Ontology Web Language*). É uma

[1] Uniform Resource Identifier é um termo técnico traduzido para a língua portuguesa como um "identificador uniforme de recurso", é uma cadeia de caracteres compacta usada para identificar ou denominar um recurso na Internet.

linguagem utilizada para definir e instanciar ontologias na web através de declarações de classes e seus relacionamentos, além de suas propriedades, regras e demais restrições. A OWL é baseada em lógica descritiva que utiliza um mecanismo de inferência para verificação de consistência e cálculo automático da hierarquia das classes envolvidas. Ela pode ser vista como uma extensão das linguagens RDF e RDF-S. Já o RDF é uma sintaxe padrão para representar um grafo dirigido em XML. Nesse âmbito, é possível construir um conjunto de dados, que se abertos, podem colaborar entre si como se a internet fosse um grande banco de dados. Muitas vezes chamado de Linked Open Data (LOD), esta virtual reunião de repositórios instrumentaliza as tecnologias da web semântica para publicar dados estruturados na web e estabelecer links entre fontes de forma interoperável.

2.2. O Modelo Conceitual Records in Contexts (RiC-CM)

O RiC-CM é resultado de mudanças significativas no domínio dos arquivos, sugeridas pelo *Expert Group for Archive Description* do Conselho Internacional de Arquivos (EGAD/ICA), após consulta pública à comunidade internacional de arquivistas (ICA/EGAD, 2023). É um modelo desenvolvido para também fundamentar a criação da ontologia RiC-O, com a proposta de substituir os padrões de descrição de arquivos vigentes como o ISAD (G), ISAAR(CPF), ISDF e ISDIAH (Miranda, 2018). Essa nova proposta possui seus fundamentos teórico-metodológicos na arquivologia e ciência da computação. Ela cria uma nova noção de "contextos" e de uma abordagem "multidimensional" no modelo conceitual RiC-CM e na ontologia RiC-O. Essa abordagem em grafos amplia as possibilidades de conexões e suas relações em um nível de contexto que pode ser estabelecido com vocabulários de uma forma não estritamente hierarquizada.

3. Linked Open Archives - LOA

A arquitetura proposta pelo protótipo de plataforma do *Linked Open Archives* (LOA) segue as etapas apresentadas na Figura 1. A partir dos arquivos disponibilizados pela plataforma, o processo de inclusão se inicia. Uma etapa de tratamento e transcrição paleográfica é realizada em um passo anterior à descrição no sistema LOA. A transcrição visa a reescrita de um manuscrito ou documento antigo para o ambiente digital, preservando com fidelidade o original. O software *Transkribus*[2] é utilizado fora do sistema nessa tarefa para reconhecer textos automaticamente validados por colaboradores dessa área. Em seguida, os registros são descritos, também pelos colaboradores, conforme o padrão RiC no software Omeka-S[3] que forma a base do sistema LOA. O Omeka se utiliza de uma base de dados relacional aonde a partir do vocabulário definido é possível manipular os objetos digitais de um acervo e descrevê-los facilmente. Uma vez que os dados estejam no banco de dados do sistema Omeka, neste caso MySQL, se faz necessário realizar um mapeamento e transformações em SPARQL para que esses dados sejam enviados para um banco de dados de triplos (*RDF Store*), nesse caso, o GraphDB[4] que também é o responsável pela criação do grafo

[2] https://www.transkribus.org/
[3] https://omeka.org/s/
[4] https://graphdb.ontotext.com/

do conhecimento onde se baseiam as consultas do LOA. O processo de leitura da base relacional e escrita no banco de dados em grafo, deve ser realizada sempre que houver atualizações nos registros do Omeka. Uma página web é então utilizada para interação com os usuários, permitindo que se realize pesquisas em registros que atendam as classes principais da ontologia RiC-O (*Agent, Event, Date, Place, RecordSource*). De forma ampliada, o usuário pode realizar uma navegação pelo grafo interativo gerado a partir da sua consulta ou pela navegação inicial. Complementarmente, o usuário pode realizar consultas SPARQL mesmo sem ter conhecimento na linguagem, através da ferramenta *Sparnatural*[5] agregada à página resultando em uma busca mais intuitiva. Essa ferramenta constrói, intuitivamente, as consultas desejadas pelo usuário, e deve ser adaptada às principais classes e propriedades existentes na ontologia base do RiC-O.

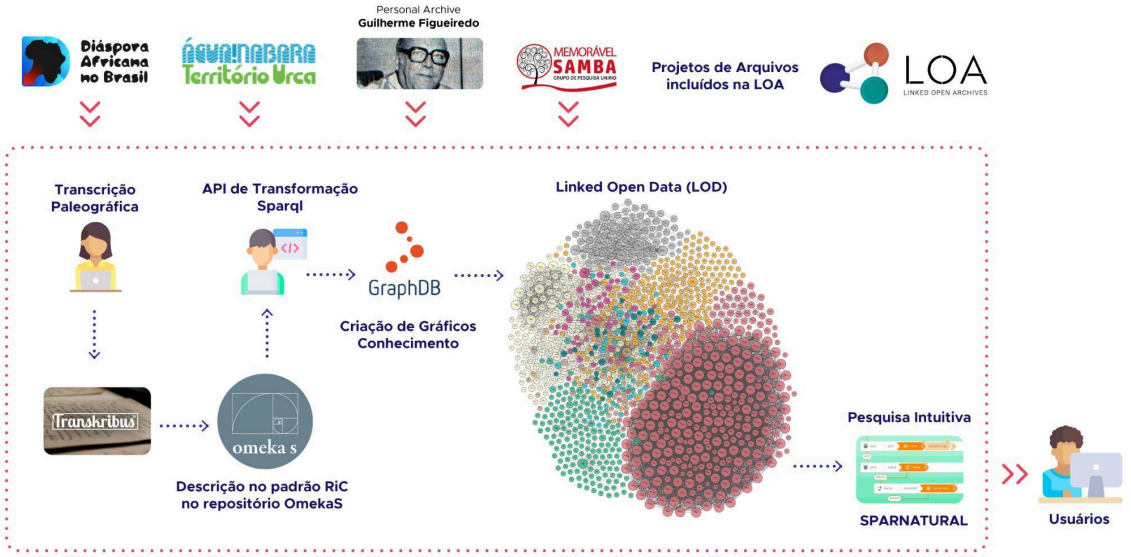

Figura 1 - Arquitetura do *Linked Open Archives* (LOA).

4. Interação em Grafos de Conhecimento do LOA

Atualmente a página inicial do *Linked Open Archives* (LOA) apresenta um grafo de algumas entidades, atributos e relações inferidas dos documentos do projeto Diáspora Africana no Brasil. Esse acervo conta com um conjunto documental da série Escravidão do Arquivo Geral da Cidade do Rio de Janeiro (AGCRJ). Esta versão inicial do grafo de conhecimento conta com algumas melhorias propostas para mitigar desafios já conhecidos, por exemplo, na visualização de grafos densos resultantes (Li et al., 2023). A Figura 2 apresenta um exemplo de trechos do grafo com algumas indicações das modificações realizadas, entre elas destacam-se:

1. **Uso do tamanho na codificação da importância dos nós**: os nós assumem um tamanho segundo o grau de arestas, tornando os nós de maior grau maiores visando destacar informações vitais ao leitor. Por exemplo, na Figura 2, temos a língua portuguesa referenciada por um nó maior que os outros, pois muitos itens descritos se relacionam a essa língua.

[5] https://sparnatural.eu

2. **Posicionamento e visibilidade dos labels**: os textos que nomeiam os nós são mantidos acima de cada item, mantendo também uma cor suave de bom contraste, na maioria dos casos o branco. Os labels de texto aparecem somente em nós de maior grau evitando que todos os nós tenham textos e o grafo fique poluído. Todos os nós e arestas mantêm informações sobre si em um elemento de dica visual (*tooltip*) aparecendo sobre o cursor do mouse.
3. **Destaque do nó e suas relações**: os nós interconectados são realçados a partir do repouso do cursor do mouse (*mouse over*), deixando os demais com maior transparência e assim facilitando a visualização. Cada item também foi mantido sem contornos nos nós;
4. **Uso de cores por projeto e elementos externos**: As cores representam os projetos inclusos na plataforma. Itens descritos pelo projeto Diáspora Africana no Brasil assumem as mesmas cores, neste caso vermelho, enquanto itens de contexto ou informações adicionais assumem outras cores, por exemplo, o branco, em links da wikipédia ou wikidata sobre a língua portuguesa. Além disso, as arestas foram mantidas com cores mais suaves para diminuir o impacto no excesso de informação com cores saturadas;

Figura 2 - Exemplos de interações na visualização do grafo do *Linked Open Archives*.

Para além das melhorias em nós e arestas, a composição do layout geral do grafo se utiliza do desenho direcionado por força (Bannister et al., 2013). Uma técnica da área de visualização de dados que conta com layout baseado na simulação de forças físicas de atração e repulsão para produção de gráficos mais espaçados e menos densos. Essa técnica também é utilizada em outros softwares de visualização de grafos de redes como o Gephi[6]. Além disso, há outros recursos interativos que facilitam a busca e detalhe das informações dispostas no grafo. Ao clicar em um dos nós, o usuário é direcionado à página do Omeka-S referente ao item pesquisado. Além disso, em informações adicionais os cliques levam diretamente a páginas relacionadas como links da wikipédia ou wikidata sobre o item interligado. Essa navegabilidade é relevante para garantir a compreensão das relações dos nós do grafo, e permitir ao usuário explorar todo o conhecimento exposto nos registros relacionados, e que estão descritos com base na ontologia RiC-O.

5. Conclusões

Este trabalho apresentou a plataforma *Linked Open Archives* (LOA) com sua arquitetura para criação de acervos interconectados por meio de tecnologias da web semântica. Da mesma forma apresentamos o grafo de conhecimento interativo que contribui para visualização e busca das diversas entidades e relações expostas pelos documentos da plataforma. Esta pesquisa se apresenta como um relato prático e experimental que contribui para demonstrar as facilidades e desafios no uso de grafos de conhecimentos na interação humano-dados. Como trabalho futuro, além das sugestões de interação já proposta pelo LOA, pretende-se investigar meios para outros desafios de visualização em grafos como filtrar, recolher ou expandir áreas durante a exploração do grafo. Além do uso de outros formatos de visualizações como tabelas em conjunto ao grafo para fornecer mais contexto na própria plataforma ao invés de direcionamento para ambientes externos (Li et al., 2023). Da mesma forma, pretende-se avançar em testes com usuários em cenários reais para atestar a usabilidade e intuitiva da busca e visualização usando os recursos do LOA.

Referências

Bannister, M. J., Eppstein, D., Goodrich, M. T., & Trott, L. (2013). Force-directed graph drawing using social gravity and scaling. In Graph Drawing: 20th International Symposium, GD 2012, Redmond, WA, USA, September 19-21, 2012, Revised Selected Papers 20 (pp. 414-425). Springer Berlin Heidelberg.

Berners-Lee, T. Linked Data. [s.l: s.n.]. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 02 set. 2024.

Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. Scientific American, v. 284, n. 5, p. 34–43, 2001.

ICA/EGAD, International Council on Archives. (2023). Records in Contexts Conceptual Model. [s.l: s.n.]. Disponível em: <https://www.ica.org/resource/records-in-contexts-conceptual-model/>. Acesso em: 02 set. 2024.

[6] https://gephi.org/

Li, H., Appleby, G., Brumar, C. D., Chang, R., & Suh, A. (2023). Knowledge graphs in practice: characterizing their users, challenges, and visualization opportunities. IEEE Transactions on Visualization and Computer Graphics.

Maciel, C., Guzman, I. R., Berardi, R. C. G., Rodriguez-Rodriguez, N., Salgado, L., Frigo, L. B., Branisa, B. & Jiménez, E. (2024). An Open Data Platform to Advance Gender Equality in STEM in Latin America. Communications of the ACM, 67(8), 90-92.

Miranda, J. M. (2018). Records in Contexts (RiC): Análise da sua aplicação em arquivos, à luz das tecnologias Linked Open Data (LOD). p. 1–26, 2018.

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. Queue, 17(2), 48-75.

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review, 56(11), 13071-13102.

Sabou, M., Simperl, E., Blomqvist, E., Groth, P., Kirrane, S., de Melo, G., Mons, B., Paulheim, H., Pintscher, L., Presutti, V., Sequeda J. F. & Shimizu, C. M. (2016). 3.24 Human and Social Factors in Knowledge Graphs. Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web, 39(4), 100.