Plan Overview

A Data Management Plan created using DMPTool

Title: EvOLoD: Evolução de Dados Interconectados na Web Semântica

Creator: Julio Cesar dos Reis

Affiliation: State University of Campinas (unicamp.br)

Data Manager: André Regino, Andressa Santos, Enio Monteiro, Rodrigo Bonacin

Funder: São Paulo Research Foundation (fapesp.br)

Funding opportunity number: 59653

Template: UNICAMP-GENERICO: Aplicável a todas as áreas

Project abstract:

A possibilidade de interconexão entre diferentes repositórios de dados estruturados pode alterar significativamente a maneira como conhecimento é produzido e consumido na Web. Para esse fim, ontologias visam representar a semântica em sistemas computacionais. Elas consistem de uma estrutura sintática que modela os conceitos de um domínio do conhecimento, e servem como schemas que organizam dados expressando instâncias de conceitos segundo propriedades lógicas. Na Web Semântica, ontologias apoiam a publicação e interconexão explícita de elemento de dados estruturados e definidos semanticamente a partir de diferentes fontes. Além do estabelecimento de correspondências entre instância de conceitos (i.e., interconexão instância-instância), ontologias são referências para a geração de anotações semânticas (i.e., ligação instância-documento). Anotações consistem em relações explícitas semanticamente denotadas a partir de um conceito da ontologia para um elemento textual de um documento Web. Isso permite a interpretação semântica de dados e das informações anotadas, e o consequente raciocínio lógico para suportar consultas sofisticadas. No entanto, a Web é dinâmica e os dados estruturados definidos tendem a evoluir através de novas versões de bases de conhecimento que são atualizadas regularmente. Esse cenário gera desafios complexos de investigação pois a definição das propriedades de conceitos e suas instâncias tendem a sofrer modificações (e.g., remoções, alterações de valores) e impactam potencialmente na qualidade de um volume enorme de anotações e interconexões existentes. Este projeto de pesquisa objetiva estudar a evolução de dados interconectados na Web Semântica. Visamos construir um framework que permite capturar e caracterizar as mudanças em repositórios de dados estruturados e implementar automaticamente adaptações sobre interconexões e anotações afetadas por essas mudanças. Primeiramente, conduziremos experimentos extensivos com dados do mundo real para entender fatores chaves que influenciam a evolução dos dados interconectados. Considerando os requisitos levantados pelos experimentos, a proposta definirá métodos originais e adequados para semi-automaticamente detectar mudanças nas bases e aplicar operações de correção sobre interconexões entre diferentes bases, e sobre anotações. O framework explora tecnologias da Web Semântica como RDF(S) e linguagens de descrição de ontologias OWL. Além da formalização dos métodos, realizaremos validações experimentais para avaliar empiricamente e

profundamente as soluções desenvolvidas em estudos de caso realísticos. Medidas de avaliação como precisão serão calculadas e analisadas para examinar a efetividade dos métodos. As contribuições esperadas envolvem novos algoritmos e ferramentas de software implementando os métodos propostos em um arcabouço completo que pode aprimorar a consistência da evolução de dados interconectados descritos semanticamente.

Start date: 10-01-2017

End date: 09-30-2021

Last modified: 01-13-2021

Copyright information:

EvOLoD: Evolução de Dados Interconectados na Web Semântica

Neste projeto, os principais dados coletados e analisados são representados em formatos de triplas em texto plano, na linguagem RDF (Resource Description Framework), com extensão .ttl, .nt e .owl.

No que diz respeito ao aspecto de manutenção de links, os dados são provenientes de bases de conhecimento/ontologias relacionadas à: (1) agricultura e saúde (Agrovoc); (2) geografia (GeoNames) e (3) enciclopédias multilingues (DBpedia e Wikidata).

Em relação à adaptação de anotações, anotações textuais exploradas são provenientes de: (1) resumos de artigos relacionadas à agricultura, alimentação e meio ambiente; (2) conteúdo de filmes em páginas Web. Os resumos são coletados em bases científicas (e.g. PubMed) enquanto os conteúdos de filmes exploram websites com conteúdos nesse domínio. Anotações são armazenadas em banco de dados para análise e investigação de sua adaptação considerando a evolução das bases de conhecimento que as geraram. A principal base explorada é o Agrovoc.

Logo, os dados coletados da pesquisa são de repositórios públicos de dados abertos na Web de dados. Nesse contexto, esta investigação produz os seguintes dados de pesquisa:

a) links atualizados entre dados de repositórios em formato RDF com base nas bases de conhecimento analisadas. Armazenaremos os links resultantes

b) anotações semânticas atualizadas com base nas anotações coletadas e que necessitam de manutenção. Anotações modificadas serão armazenadas e compartilhadas

c) algoritmos originais para a manutenção de links e anotações no contexto de Web semântica

d) código em linguagem Java de um sistema computacional dedicado a manutenção de links e anotações. O sistema permitirá uma abordagem assistida em que especialistas (gestores das bases de conhecimento) analisem os links.

e) documentação para descrição e manual de uso da ferramenta.

Esses itens serão compartilhados a partir de um arquivo zip compactado.

O padrão Dublin Core será adotado como modelo de metadados. Os metadados são padronizados expressam informações sobre os autores, título, descrição, licença para uso, data de criação, nome do programa e tags para facilitar a busca do objeto da pesquisa.

Não há necessidade de consulta ao Comitê de Ética, uma vez que os dados utilizados não se referem a seres humanos, espécies em extinção ou outros que requeiram cuidados especiais. Desta forma, os dados serão disponibilizados para uso de maneira livre sobre a licença CC0 e disponíveis para acesso público.

Não há restrições em relação ao compartilhamento dos dados usados e obtidos na pesquisa em questão, independente da pessoa que fará uso ou do objetivo pretendido. Os dados estarão disponíveis tanto em repositórios oficiais da universidade do executor (em forma de resultados de pesquisa), quanto em repositórios públicos de dados abertos (dados de entrada para a implementação da pesquisa), alvo do trabalho do responsável por este projeto. Todos os dados gerados por esta pesquisa poderão ser utilizados para fins de contribuir com outras pesquisas, desde que os autores originais sejam citados. Resultados da pesquisa serão armazenados no repositório oficial da UNICAMP e terão DOI gerado como identificador único e persistente para sua citação.

Os arquivos resultantes são da mesma natureza que os arquivos usados como entrada de dados para a condução da pesquisa: arquivos em extensão .ttl ou .nt, formatados em triplas. Os dados serão disponibilizados no formato RDF/TTL e CSV, devido a facilidade de importação e visualização em programas como Apache JENA e Excel. Esses arquivos podem ser abertos por qualquer editor de textos convencional e podem ser processados por qualquer ferramenta de leitura de arquivos RDF ou visualizadores gráficos de triplas. Esta pesquisa também gerará dados em formato .tex na confecção de documentos de relatam os resultados da investigação.

Todos os arquivos gerados serão armazenados em sistemas de armazenamento em nuvem que podem ser facilmente acessados pelo pesquisador e por terceiros, que terão permissão d

sx e efetuar cópia dos dados e modificá-las de acordo com suas necessidades. Durante a pesquisa, os dados serão mantidos no HD do laboratório e gerenciados pela equipe de trabalho. Além disso, serão utilizados os sistemas de armazenamento mantidos pela universidade, como o Google Drive, vinculado ao e-mail institucional do aluno. Ao final, os dados gerados serão disponibilizados no repositório oficial - REDU UNICAMP - seguindo as políticas de preservação da mesma instituição. Os dados de entrada, saída, resultados parciais e quaisquer outros produtos derivados da pesquisa, código fonte, serão armazenados no serviço acima mencionado além de serem disponibilizados na plataforma online Gitlab. Tal plataforma é vinculada a conta institucional do pesquisador e está sob resguardo da universidade, onde são mantidas e preservadas pela equipe de TI. Adicionalmente, as publicações deste projeto são catalogadas no Repositório da Produção Científica e Intelectual da Unicamp.