

Uma Solução Semi-Automática para Extração, Transformação e Carga de Dados Abertos Conectados

Sérgio Souza Costa, Mateus Vitor D. Sousa, Micael Lopes da Silva, Eddy Cândia de Oliveira, José Victor M. Guimarães



WIDaT
2019



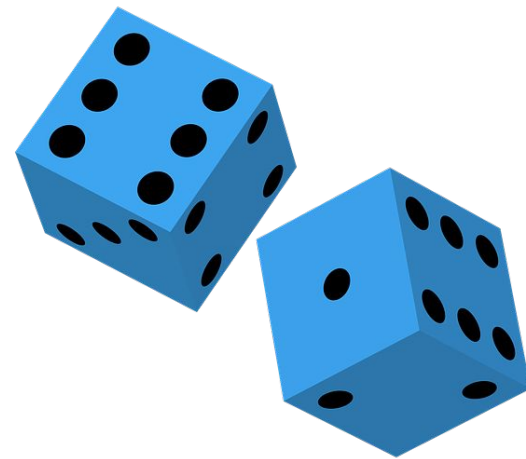
III Workshop de Informação Dados e Tecnologia - WIDaT
Curso de Bacharelado Interdisciplinar em Ciência e Tecnologia - BICT
Universidade Federal do Maranhão - UFMA

Agenda



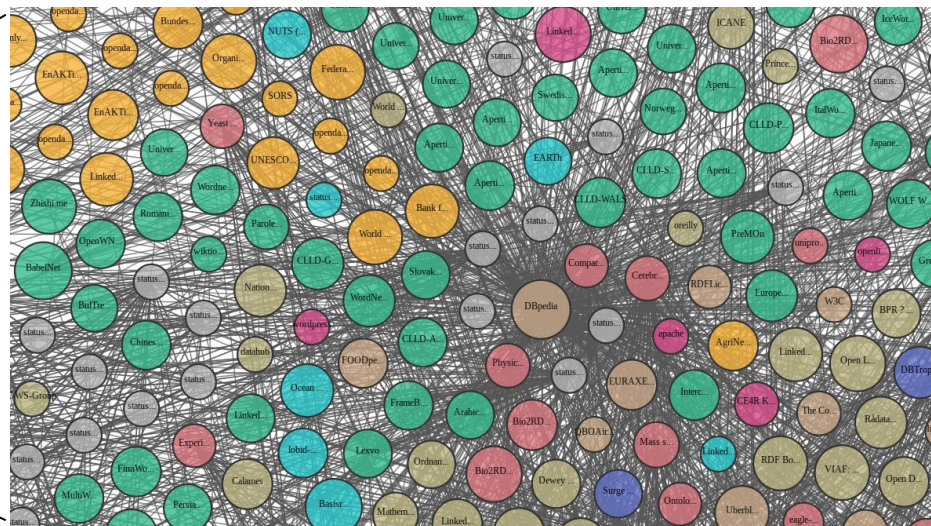
1. Introdução
2. Trabalhos Relacionados
3. Estudo de Caso
4. Metodologia
5. Arquitetura
6. Resultado
7. Conclusão
8. Trabalhos Futuros
9. Referências

- Dados e Informações;
- Acesso aos dados públicos já eram previstos pela constituição de 1988 e reforçado pela Lei Nacional de Acesso à Informação (nº 12.527/2011).

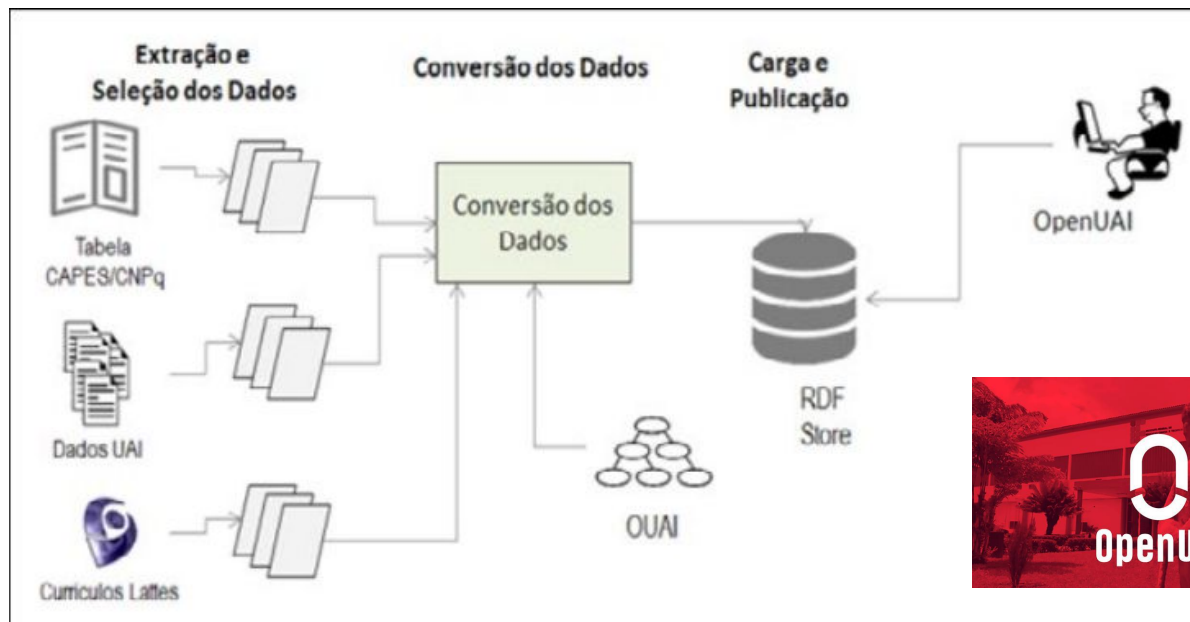


- Necessidade de maior comunicação e transparência;
- Surgimento do paradigma de Dados Abertos;
- Democratização de dados e o incentivo a comunidade na geração de novas soluções acerca destes dados.

- Paradigmas de Dados Conectados proposto por Tim Berners-Lee em 2011;
- Surgimento de novos problemas e adequação ao recente cenário da web;
- Dados Conectados, em suma, é um conjunto de boas práticas para a estruturação, interligação e armazenamento dos dados;
- Utilização de URI's;



- Utilização de Vocabulários e URI's;
- No contexto das universidades identificamos alguns trabalhos relacionados como em Pantoja (2013), D'Aquin et. al (2014) Kessler e Kauppinen (2012), Alencar, Xavier e Souza (2018).



Arquitetura da OpenUAI do Instituto Federal da Paraíba - IFPB

Diese Website verwendet Cookies. Wenn Sie die Website weiter nutzen, gehen wir davon aus, dass Sie hiermit einverstanden sind.

[Datenschutzhinweis](#)

OK



[Data](#) [Research](#) [Team](#) [About](#)

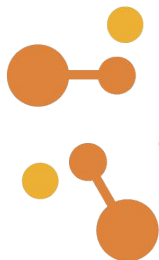
Welcome...

...to the University of Münster's Open Data initiative!



Linked Open Data University of Munster (Alemanha).

Open Knowledge Graph



Linked Data from The OU

We interlink and expose data available from open institutional repositories of [The Open University](#) and make it available for reuse.

The data can be searched via a [keyword-based interface](#), using [well-known identifiers](#), or queried using the [SPARQL endpoint](#).

Our datasets relate the publications, qualifications, courses and Audio/Video material produced by The Open University, as well as the people involved in making them.

All these data are available through standard formats (RDF and SPARQL) and are (in most cases) available under an open license.

Search // Lookup

 Keyword	eg: linked data	
 OU	eg: ed4565	
 OU Course code	eg: M366	
 OU Qualification code	eg: Q67, D37	
 YouTube Video Code	eg: 0q-Y9U46Qc4	
 ORO identifier	eg: 21166	

Open Knowledge Graph, projeto da Open University (Reino Unido).

- Universidade Federal do Maranhão distribuída por todo estado maranhense;
- Conjunto de portais públicos com os dados disponíveis;
- Portal de dados abertos oficiais;

- **ETL (Extract, Transform and Load)**, conceito aplicado em Data Warehouse que apoiou o esquema arquitetural.

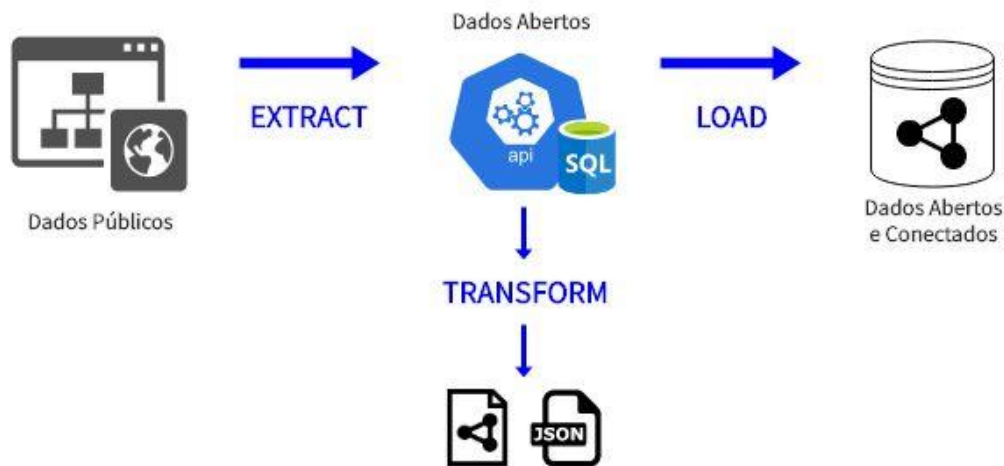


Figura: Esquema simplificado do ciclo de vida dos dados.

Arquitetura

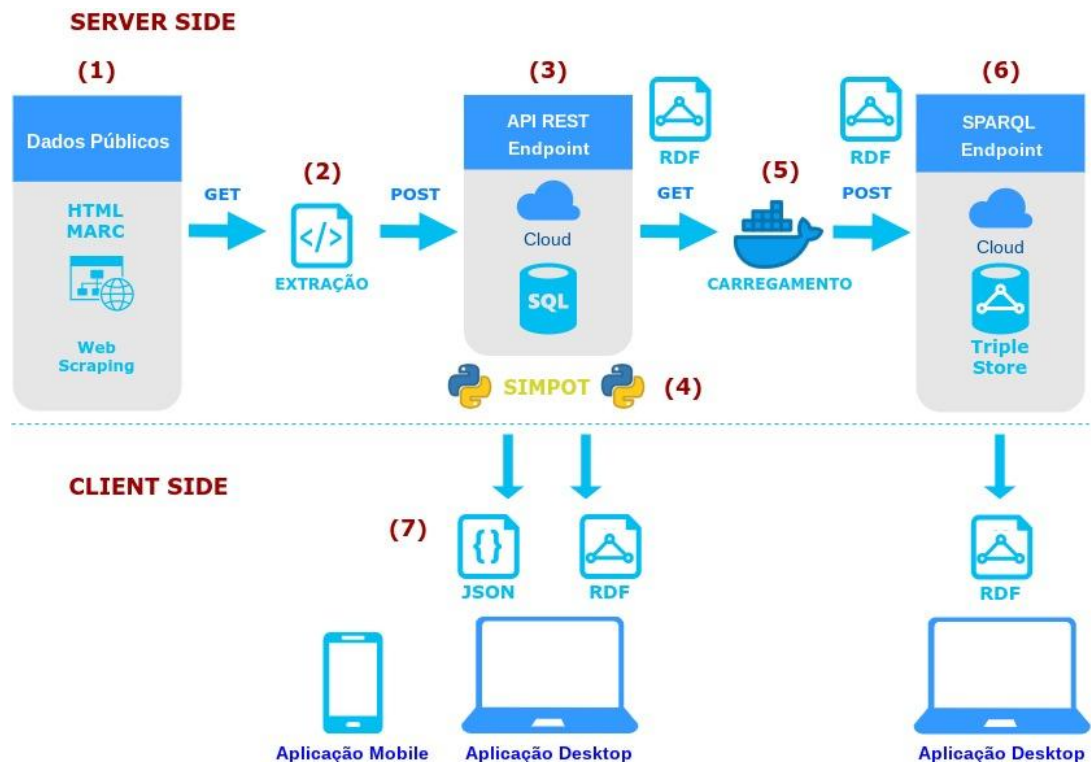


Figura: Esquema completo da arquitetura da aplicação.

- Web Scraping dos portais públicos da universidade (*dados públicos*);
- Adequação dos dados para o paradigma de **Dados Abertos**;
- Utilização do padrão REST. Criação de **API's REST** para recuperação dos dados por meio de requisições HTTP.

- Conversão dos dados de formato aberto para dados conectados;
- Transformação para RDF - Resource Description Framework;
- SIMPOT - Simple Object-Triple;
- Módulo desenvolvido em python e inspirado no SQLAlchemy;

```
base = "https://sigaa.ufma.br/sigaa/public"
class Docente:
    nome = FOAF.name
    email = FOAF.mbox

    @RdfsClass(FOAF.Person,
        base + "/docente/portal.jsf?siape=")

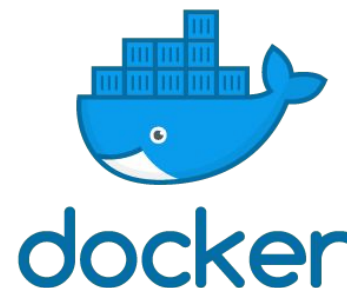
    @BNamespace("dc", DC)
    @BNamespace("foaf", FOAF)

    def __init__(self, nome, email, mat):
        self.nome = Literal(nome)
        self.email = URIRef(email)
        #convenção para gerar o URI
        self._id = mat

d=Docente("Mateus Vitor", "mateus@gmail.com", 685)
print (graph (d)) # imprime o grafo
```

Listagem: Exemplo da criação de um dado RDF utilizando a SIMPOT.

- Carregamento dos dados RDF para um Triple Store;
- Os dados são baixados e carregados de uma só vez;
- Docker encarregado de deixar a aplicação operacional.



- Duas alternativas para consumo de dados na aplicação;
- Dados Abertos e os REST Endpoints;
- 45.141 registros;
- Dados Abertos e Conectados com um SPARQL Endpoint.
- 169.228 triplas;

- Consultas SPARQL;
- Consumindo aquilo que é realmente necessário por meio de um único Endpoint;
- Necessidade de entender a estrutura da entidade para um melhor aproveitamento do paradigma.



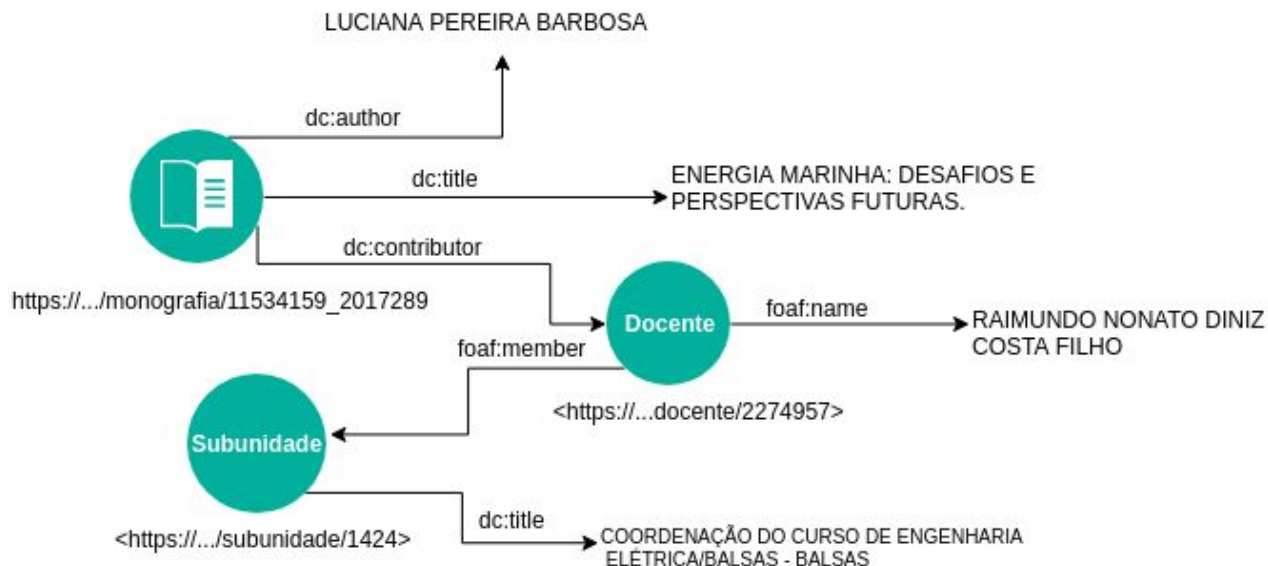


Figura: Esquema estrutural das entidades relacionando-se.

```
SELECT
?mono ?autor ?nome_curso ?nome_orientador
WHERE {
  ?s rdf:type bibo:Thesis.
  ?s dc:creator ?autor.
  ?s dc:title ?mono.
  ?s dc:publisher ?curso.
  ?s dc:contributor ?orientador.
  ?curso dc:title ?nome_curso.
  ?orientador foaf:name ?nome_orientador.
}
```

Listagem: Exemplo de consulta
SPARQL de dados de monografias.

Resultado



Figura: Nuvem de palavras gerada a partir dos dados de monografias.

- Atualização e expansão dos dados abertos e conectados ainda é um desafio;
- Com a arquitetura proposta o processo de atualização e expansão se tornou mais simples;
- A metodologia se mostrou replicável para outras instituições;
- Não foram feitos testes rigorosos de eficiência;
- Destaque para as tecnologias SIMPOT, Jena Fuseki, Docker e Heroku.

- Avaliação e testes com outros triples stores, por exemplo, o VIRTUOSO;
- Análise mais profunda para melhoria semântica na utilização de vocabulários já existentes;
- Implementar novas aplicações que aproveitem esta base de dados gerada.

1. **Scraper:** https://github.com/inovacampus/ufma_scrapper
2. **Live API:** <https://github.com/inovacampus/ufma-live-api>
3. **Database API:** <https://github.com/inovacampus/ufma-api>
4. **SIMPOT:** <https://github.com/inovacampus/simpot>
5. **LODUFMA:** <https://github.com/inovacampus/linked-open-data-ufma>

ALENCAR, A.; XAVIER, D.; SOUZA, D. Publicação e consumo de dados abertos conectados acadêmicos. Revista Principia, 2018.

BERNERS-LEE, T. Linked Data. 2006. Disponível em: <<https://www.w3.org/DesignIssues/LinkedData.html>>.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, p. 205–227, 2009.

DAGA, E.; D'AQUIN, M.; ; ADAMOU, A.; BROWN, S. The open university linked data – data. open.ac.uk. Semantic Web, v. 7, n. 2, p. 183–191, 2015. Disponível em: <<http://www.semantic-web-journal.net/system/files/swj973.pdf>>.

KESSLER, C.; KAUPPINEN, T. Linked open data university of münster–infrastructure and applications. In: SPRINGER. Extended Semantic Web Conference. 2012. p. 447–451. Disponível em: <<http://kauppinen.net/tomi/lodum-eswc-2012.pdf>>.

LEDVINKA, M.; KREMEN, P. A comparison of object-triple mapping frameworks. Semantic Web. 2018

PANTOJA, J. Linked Open Data at the UPF. 2013. Disponível em: <<http://data.upf.edu/upf/docs/2013/jorgepm/lodatupf.pdf>>.

Uma Solução Semi-Automática para Extração, Transformação e Carga de Dados Abertos Conectados

Sérgio Souza Costa, Mateus Vitor D. Sousa, Micael Lopes da Silva, Eddy Cândido de Oliveira, José Victor M. Guimarães



WIDaT
2019



III Workshop de Informação Dados e Tecnologia - WIDaT
Curso de Bacharelado Interdisciplinar em Ciência e Tecnologia - BICT
Universidade Federal do Maranhão - UFMA