CENTRO DE PESQUISA E FORMAÇÃO
SESC

Dados e Repositórios Culturais

AULA 2

Web Semântica e Repositórios

Sobre o Curso

CENTRO DE PESQUISA E FORMAÇÃO SESC

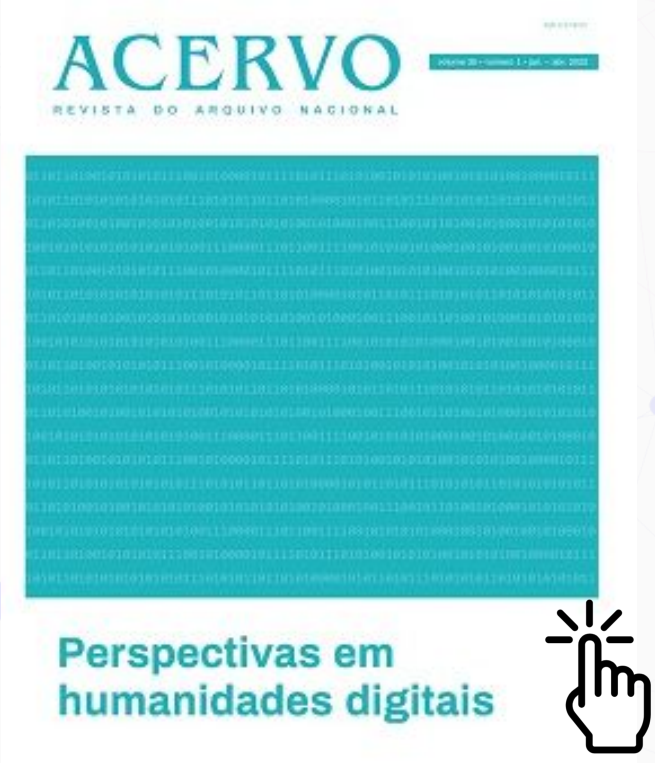
Slides, Artigos, Materiais...

Sobre o Curso

CENTRO DE PESQUISA E FORMAÇÃO SESC

CO-CONSTRUÇÃO

Quem leu?



Identificando dados de pesquisa nas humanidades

Márcia Cavalcanti

CAVALCANTI, M. Identificando dados de pesquisa nas humanidades. Acervo, [S. l.], v. 35, n. 1, p. 1–18, 2022. Disponível em: https://revista.an.gov.br/index.php/revistaacervo/article/view/1775.

Quem leu?

Plano de Gestão de Dados (PGD):

Documento que descreve como os dados de uma pesquisa serão coletados, organizados, armazenados, preservados e compartilhados.

Por que é importante?

- Facilita a reprodutibilidade da pesquisa
- Garante conformidade com políticas institucionais e de fomento
- Melhora a visibilidade e o impacto dos dados gerados

Elementos típicos de um PGD:

- Tipos de dados e formatos
- Padrões de metadados
- Direitos autorais e ética
- Armazenamento e backup
- Acesso, reuso e preservação de longo prazo

https://fapesp.br/gestaodedados

Quem leu?

CENTRO DE PESQUISA E FORMAÇÃO SESC

"A gestão de dados científicos cobre todo o chamado "ciclo de vida" dos dados, ou seja, desde a sua coleta até o armazenamento de longo prazo, passando por uma série de processamentos de limpeza, curadoria, anotação, indexação e transformação. Grande parte da pesquisa científica de hoje exige algum tipo de análise e processamento de dados. Com isto, o planejamento da gestão dos dados utilizados e gerados em uma pesquisa passou a fazer parte integral da metodologia científica, sendo, inclusive, considerado como um dos itens necessários de boas práticas de pesquisa. **(Medeiros, 2018)**

Quem leu?

Dados de pesquisa em formato não digital[7]	Dados de pesquisa em formato digital
Notas de campo escritas em cadernos de pesquisa Fotos (impressas) Mapas feitos à mão Transcrições de trocas verbais e comportamentais (impressas) Questionários (impressos) Experimental Observacional Modelos Transcrições (impressas) Testes padronizados (impressos) Métodos usados para geração de dados Dados de biomarcadores coletados a partir de medições físicas Amostras biológicas dos entrevistados	Notas de campo escritas em cadernos de pesquisa digitais ou digitalizadas posteriormente Fotos digitais ou digitalizadas posteriormente Mapas digitais ou digitalizados posteriormente Dados quantitativos Dados que foram analisados usando software qualitativo de análise de dados Medidas de resultado: tempos de resposta; taxas de erro Gravações em áudio ou vídeo de respostas verbais ou motoras Medidas on-line: potenciais relacionados ao evento; rastreamento ocular; rastreamento do *mouse* Imagem cerebral funcional Gravações de vídeo ou áudio Transcrições de trocas verbais e comportamentais Questionários Dados baseados em computador: tela sensível ao toque; rastreamento ocular; tutores Arquivos simples baseados em texto para análise estatística Conjuntos de dados vinculados Dados públicos

Web Semântica

A Web Semântica é uma extensão da internet atual onde os dados ganham sentido, se conectam com lógica e podem ser interpretados por máquinas com inteligência.

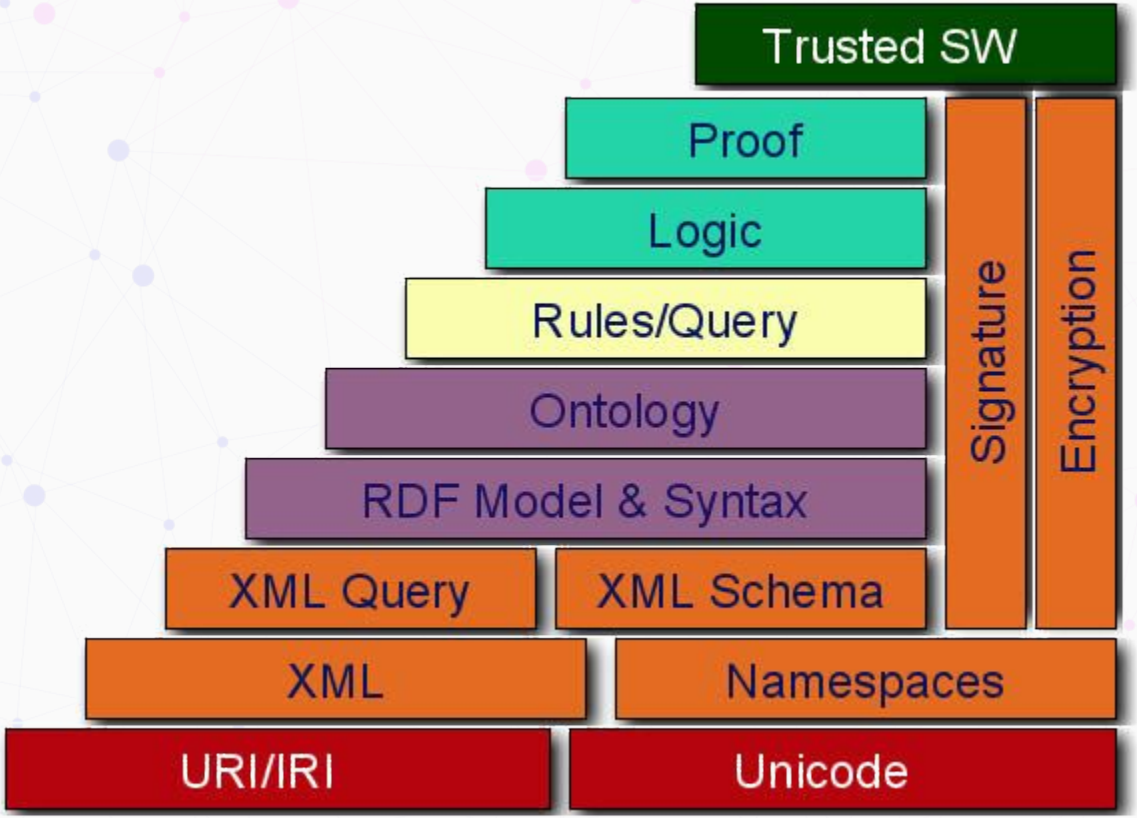
Enquanto a web tradicional mostra palavras para humanos, a Web Semântica organiza **conceitos, relações e contextos** de forma que computadores também consigam compreender, cruzar e raciocinar sobre o conhecimento.

Ela transforma a internet em um espaço de pensamento colaborativo, onde acervos culturais, obras artísticas, movimentos históricos, línguas e práticas sociais podem:

- ter seus significados formalmente descritos,
- se conectar a outros dados no mundo,
- responder perguntas complexas,
- explicar suas decisões,
- e proteger a autoria, a memória e o direito à informação segura.

É como ensinar a internet a pensar com cuidado e com contexto.

Web Semântica



URI/IRI

URI (Identificador Uniforme de Recursos)

IRI (Identificador Internacionalizado de Recursos)

São como “endereços digitais” que identificam de forma única qualquer coisa na web: uma pessoa, um conceito, um documento, um lugar etc.

Isso permite que essas informações sejam referenciadas de forma inequívoca na internet, em qualquer idioma.

Exemplos de URIs:

URN (nome):

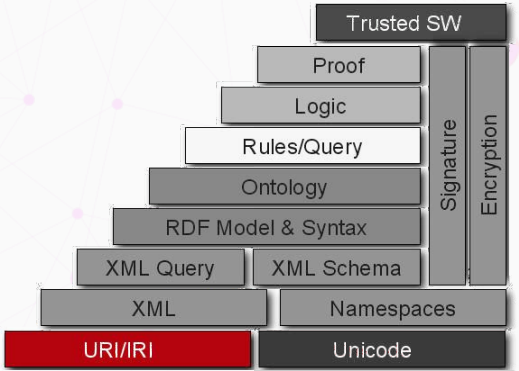
urn:isbn:0-486-27557-4

identifica uma edição específica de um livro, mas não indica como acessá-lo

URL (localizador):

https://www.exemplo.com/

identifica e localiza a página principal do site exemplo.com



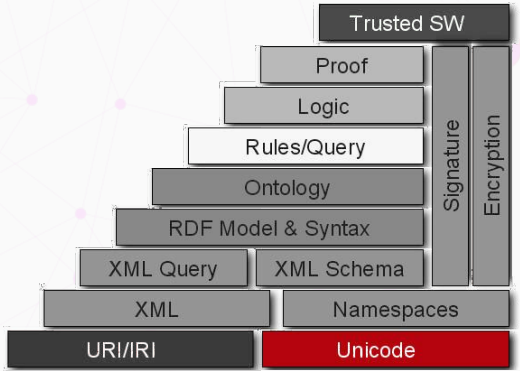
Unicode

É o sistema que permite representar todos os caracteres das línguas humanas (nomes, textos, títulos e obras em português, árabe, japonês, guarani, etc.) em formato digital.

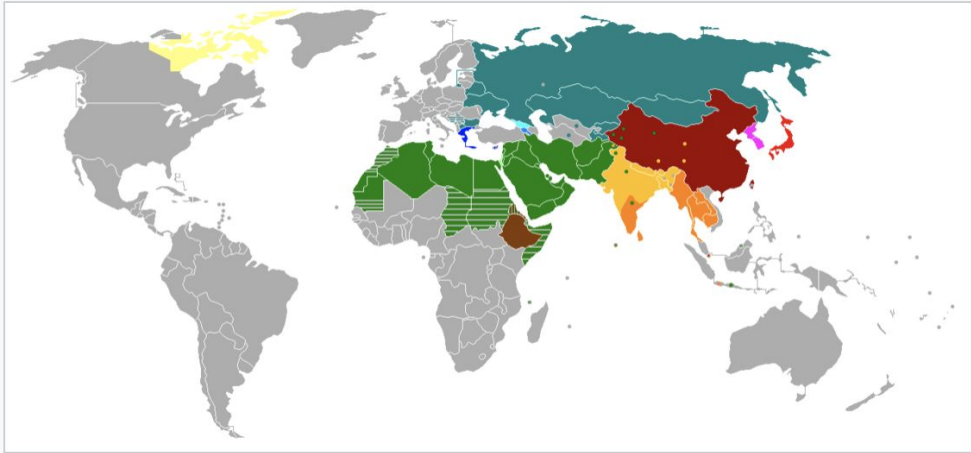
Definições-chave:

Padrão universal de codificação de caracteres.

É uma das camadas mais invisíveis e fundamentais da Web Semântica, e entender suas aplicações ajuda a perceber como ele viabiliza o mundo digital multilíngue e multicultural em que vivemos



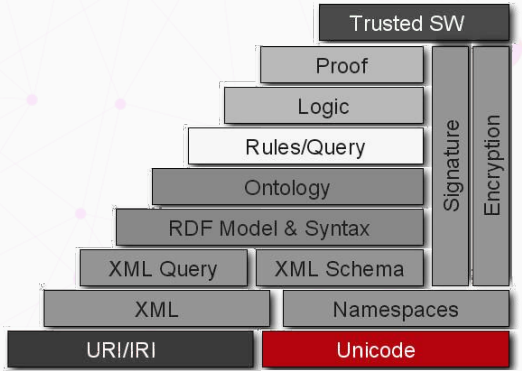
Unicode



Index of predominant national and selected regional or minority scripts

Alphabetic	[L]ogographic and [S]yllabic	Abjad	Abugida
Latin	Hanzi [L]	Arabic	North Indic
Cyrillic	Kana [S] / Kanji [L]	Hebrew	South Indic
Greek	Hanja[b] [L]		Ethiopic
Armenian			Thaana
Georgian			Canadian syllabic
Hangul[a]			

[a] Featural-alphabetic. [b] Limited.



XML

CENTRO DE PESQUISA E FORMAÇÃO SESC

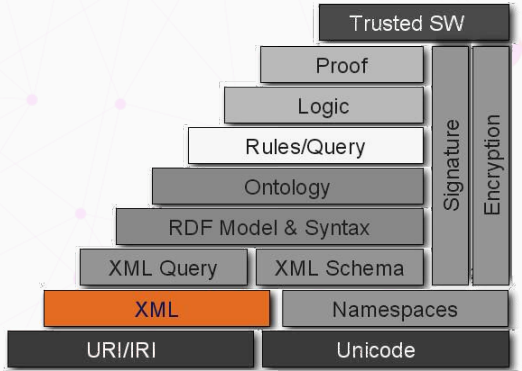
XML (eXtensible Markup Language) é uma **linguagem de marcação** que permite organizar, estruturar e transportar dados de forma legível tanto por humanos quanto por máquinas.

Não é uma linguagem de programação: é uma forma de descrever dados com significado.

```xml
<obra>
    <titulo>Marília de Dirceu</titulo>
    <autor>Tomás Antônio Gonzaga</autor>
    <ano>1792</ano>
    <genero>Poesia</genero>
    <descricao>Obra representativa do Arcadismo brasileiro.</descricao>
</obra>
```

TEI

DublinCore



Namespaces

Namespaces ajudam a evitar confusão quando usamos termos iguais de áreas diferentes.

Por exemplo, garantem que “autor” em museologia não seja confundido com “autor” em direito.

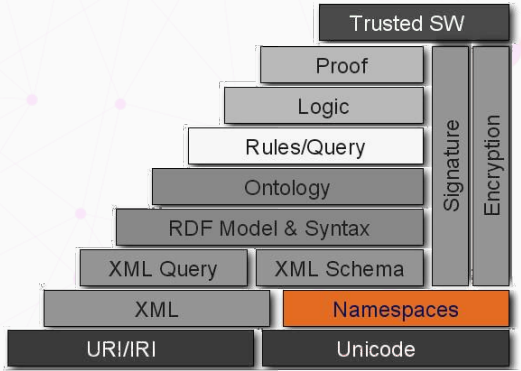
Essencial quando queremos unificar dados vindos de **fontes diferentes** sem cair em confusão de significados.

<table>
(Tabela HTML)

```
<table>
  <tr>
    <td>Apples</td>
    <td>Bananas</td>
  </tr>
</table>
```

<table>
(Móvel)

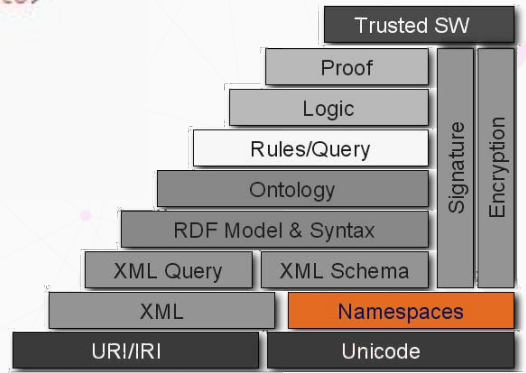
```
<table>
  <name>African Coffee Table</name>
  <width>80</width>
  <length>120</length>
</table>
```



Namespaces

```
<raiz>

  <banco>
    <nome>Banco Central do Brasil</nome>
    <agencia>001</agencia>
    <conta>12345-6</conta>
  </banco>

  <banco>
    <material>Madeira</material>
    <comprimento>150</comprimento>
    <cor>Marrom</cor>
  </banco>

</raiz>
```

```
<raiz xmlns:fin="http://exemplo.com/financeiro"
      xmlns:mob="http://exemplo.com/mobilia">

  <fin:banco>
    <fin:nome>Banco Central do Brasil</fin:nome>
    <fin:agencia>001</fin:agencia>
    <fin:conta>12345-6</fin:conta>
  </fin:banco>

  <mob:banco>
    <mob:material>Madeira</mob:material>
    <mob:comprimento>150</mob:comprimento>
    <mob:cor>Marrom</mob:cor>
  </mob:banco>

</raiz>
```



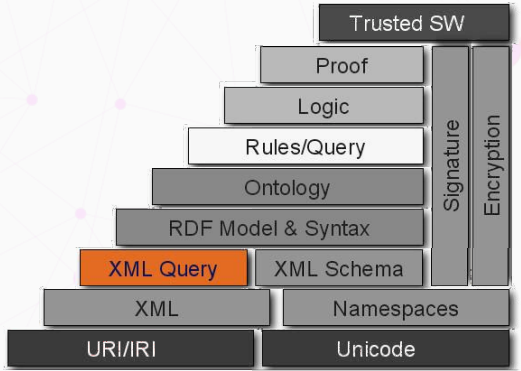
XML Queries (XQuery)

XQuery é uma linguagem usada para **fazer perguntas e extrair informações** de documentos XML, de forma inteligente e precisa.

XML por si só é apenas estrutura; o XQuery dá **vida** a esses dados, permitindo buscas, filtros e agrupamentos.

É **parecido com SQL** para bancos de dados relacionais, mas voltado para a estrutura em árvore do XML.

```
declare namespace mob = "http://exemplo.com/mobilia";

for $banco in //mob:banco
where $banco/mob:cor = "Marrom"
return $banco
```



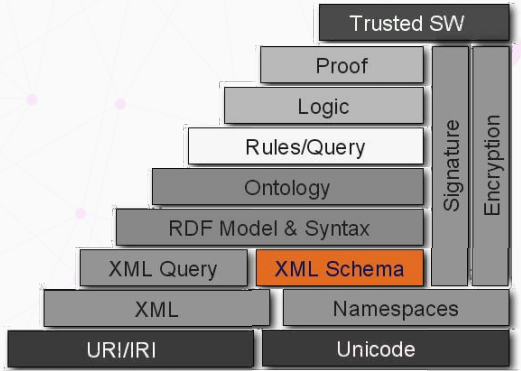
XML Schema (XSD)

CENTRO DE PESQUISA E FORMAÇÃO SESC

XML Schema é um modelo ou conjunto de regras que define como os dados em um documento XML devem estar **organizados**.

Imagine que você vai montar um **armário de arquivos**. O XML são os documentos que você coloca dentro. O Schema é o manual que diz como organizar: qual documento vai em qual prateleira, o que ele deve conter e em que formato.

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:element name="obra">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="titulo" type="xs:string"/>
        <xs:element name="autor" type="xs:string"/>
        <xs:element name="data_criacao" type="xs:date"/>
        <xs:element name="genero" type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

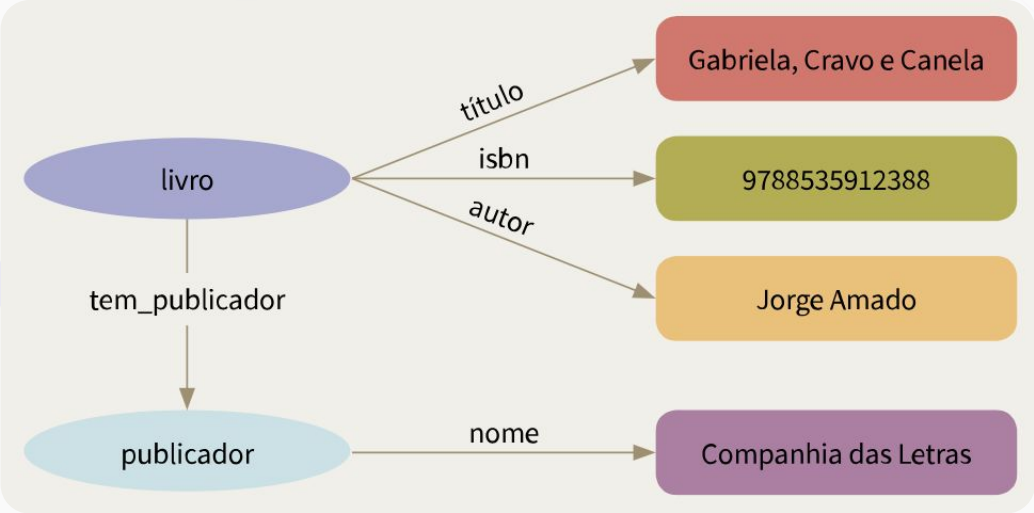


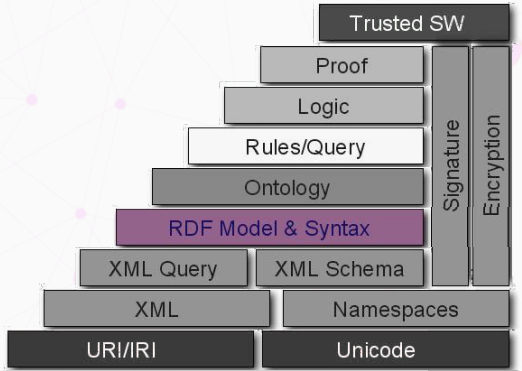
Modelo RDF e Sintaxe

RDF (Resource Description Framework) é um modelo para representar conhecimento na web de forma estruturada e relacional. RDF permite fazer **afirmações sobre recursos**. Recursos são quaisquer coisas, tanto concretas quanto abstratas. Uma determinada empresa, uma pessoa, uma página Web são considerados recursos. Um sentimento, uma cor, também são recursos.

Ele organiza a informação usando trios (triplets):

👉 **Sujeito – Predicado – Objeto**





Ontology

OWL (Web Ontology Language) é a linguagem padrão para criar ontologias.

Ajuda a formalizar o conhecimento de uma área: por exemplo, o que é um artista? Uma obra? Um curador? Como eles se relacionam?

É definida por:

👉 **Classes (ou Conceitos)**

Autor, Obra, Movimento_Artístico, Local, Gênero_Literário

👉 **Propriedades (ou Relações)**

escreveu, nasceu_em, participou_de, é_parte_de, foi_influenciado_por

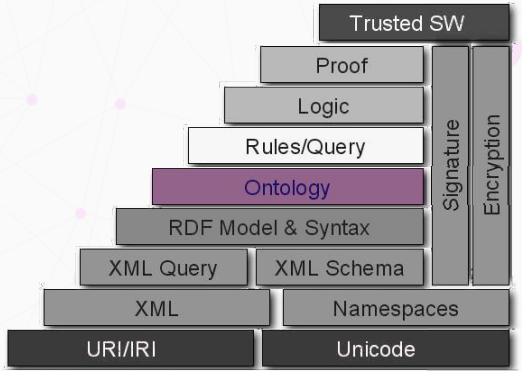
👉 **Instâncias (ou Exemplos concretos)**

Machado_de_Assis, Dom_Casmurro, Realismo, Rio_de_Janeiro

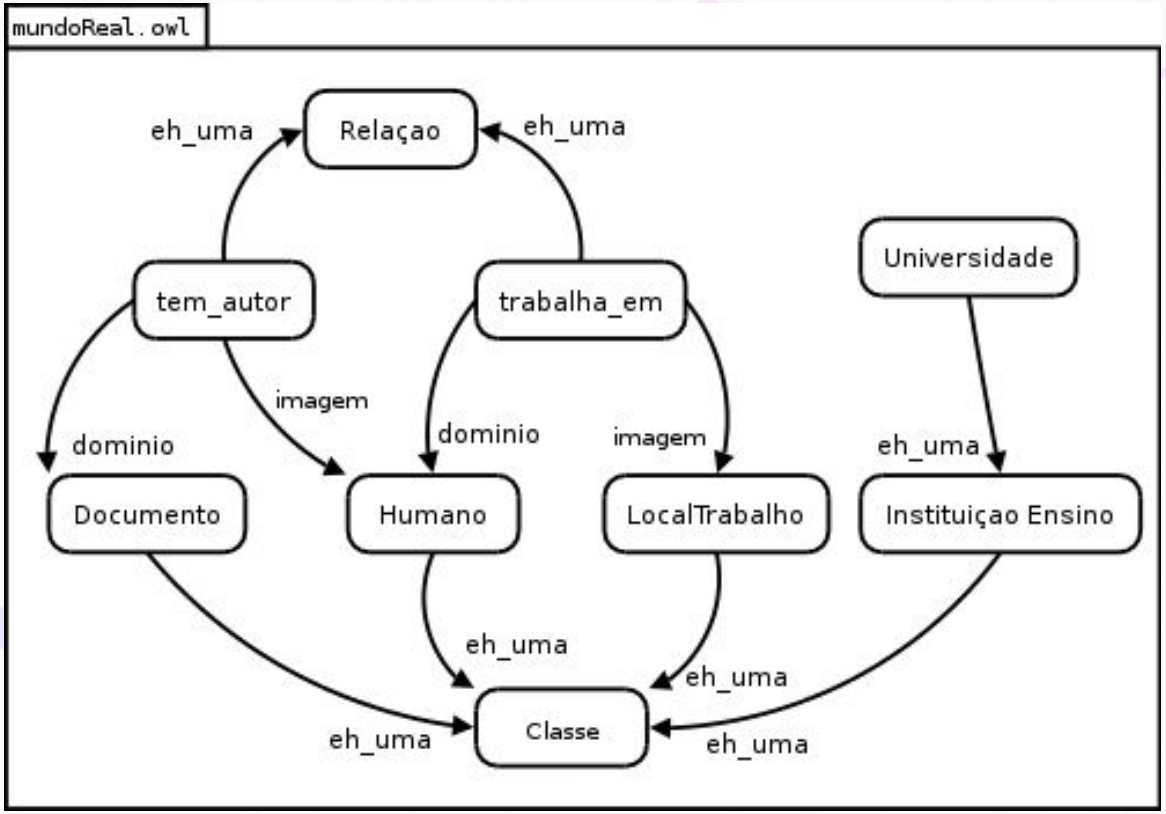
👉 **Regras e restrições**

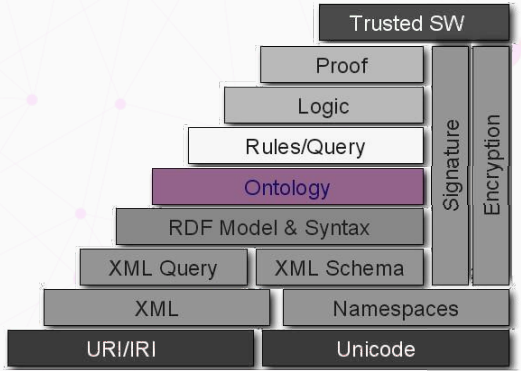
“Todo autor tem pelo menos uma obra.”

“Um movimento artístico não pode ter data posterior à morte de seu fundador.”



Ontology – OWL





Rules/Query

Esta camada permite criar regras lógicas para que os dados se comportem de maneira mais inteligente. É como ensinar o sistema a entender relações e implicações: “Se A acontece, então B é verdade”.

Regras: instruções do tipo “SE... ENTÃO...”.

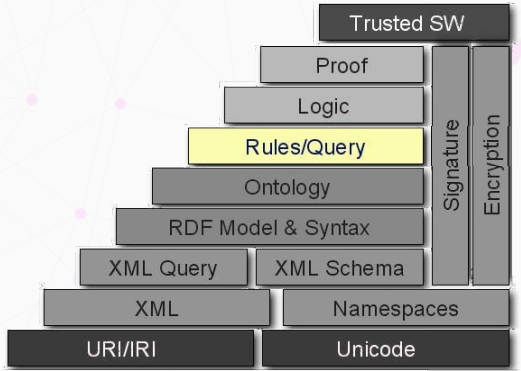
Consultas semânticas (**SPARQL**): perguntas que extraem dados com base nessas relações inteligentes.

ATRIBUIÇÃO

```
IF autor/nascimento < "1900-01-01"
AND autor/pertence_ao_movimento = Realismo
THEN autor/classe_temporal = "Século XIX"
```

CONSULTA

```
SELECT ?autora ?obra WHERE {
  ?autora rdf:type :Mulher .
  ?obra rdf:type :Romance .
  ?obra :escrita_por ?autora .
  ?obra :ano_publicacao ?ano .
  FILTER (?ano >= 1930 && ?ano <= 1960)
}
```



Lógica

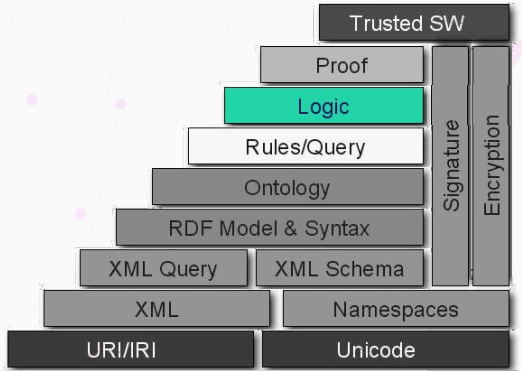
A Web Semântica visa transformar a Web de um repositório de dados planos em um sistema onde é possível expressar **lógica**. Diferentemente dos sistemas tradicionais de representação do conhecimento, que priorizam eficiência computacional, a Web Semântica busca ser um sistema unificador, permitindo que a complexidade da realidade seja descrita com expressividade total. Isso significa que, embora a Web Semântica não defina um mecanismo de raciocínio específico, ela estabelece operações válidas e exige consistência para elas.

🧩 Princípios Fundamentais

1. Expressividade sobre Eficiência

2. Validação de Provas em vez de Demonstração

3. Mínimo Conjunto de Regras Universais

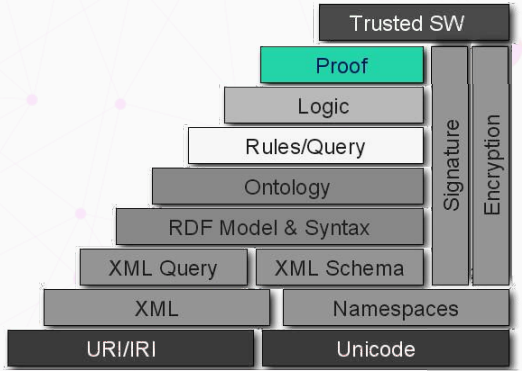


Proof

A camada de **Proof (Prova)** se refere à capacidade de um sistema semântico de explicar como chegou a uma determinada inferência ou conclusão, de maneira estruturada e verificável por outros sistemas ou por humanos.

Essa prova pode ser usada para:

- **Auditar processos de inferência**
- **Verificar coerência de dados**
- **Estabelecer confiança em contextos sensíveis**
- **Justificar recomendações culturais, educacionais, ou históricas**



Assinatura Digital

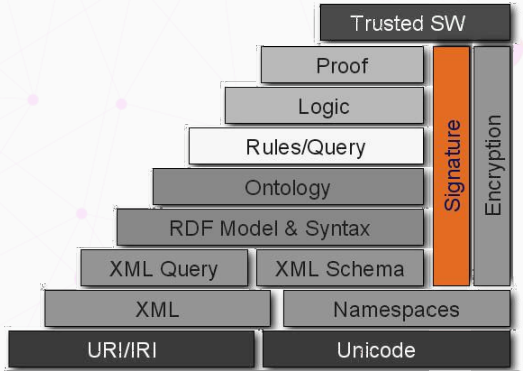
É um mecanismo criptográfico que permite autenticar a autoria de um dado digital.

Garante que o dado:

- **veio de quem diz que veio,**
- **não foi alterado no caminho,**
- **e pode ser verificado por qualquer um com a chave pública.**

Pense nela como uma assinatura de documento físico, mas:

- **infalsificável,**
- **automatizável,**
- **rastreável por sistemas semânticos.**



Criptografia

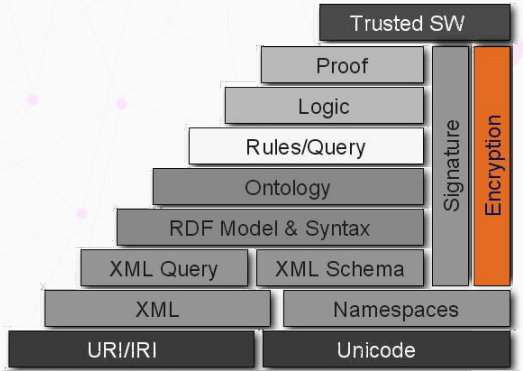
É o processo de proteger os dados contra acessos não autorizados, transformando a informação em um conteúdo codificado que só pode ser lido por quem possui a chave correta.

A criptografia garante:

- **Confidencialidade**
- **Controle de acesso**
- **Proteção contra censura ou manipulação**

Métodos de criptografia:

- **Criptografia Simétrica**
- **Criptografia Assimétrica**



Repositórios de Dados Públicos

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

órgão federal responsável pelas evidências educacionais e atua em três esferas: avaliações e exames educacionais; pesquisas estatísticas e indicadores educacionais; e gestão do conhecimento e estudos.

Instituto Brasileiro de Geografia e Estatística (SIDRA)

disponibiliza uma vasta base de dados sobre o Brasil, abrangendo diversos aspectos, como geografia, população, economia, sociedade e território.

Instituto de Pesquisa Econômica Aplicada (Ipeadata)

O Ipeadata possui bases de dados Macroeconômicos, Regionais e Sociais.

Repositórios de Dados Públicos

CENTRO DE PESQUISA E FORMAÇÃO SESC

Supremo Tribunal Federal (Corte Aberta)

visa garantir que os dados da Corte sejam disponibilizados a todos os cidadãos de maneira mais acessível, precisa, confiável e íntegra – observando-se os pilares da proteção de dados pessoais e da segurança cibernética.

Tribunal Superior Eleitoral

disponibiliza à sociedade os dados gerados ou custodiados pelo TSE, de forma a garantir o acesso a informações e aprimorar a cultura de transparência. Ele substitui o antigo Repositório de Dados Eleitorais, descontinuado em janeiro de 2022.

Banco Central do Brasil

Ampliar e aprimorar no BC a transparência ativa por meio da abertura de dados públicos, com eficiência e qualidade, de forma a contribuir para reforçar a credibilidade e o cumprimento da missão institucional do BC, bem como fomentar o controle social, o aperfeiçoamento da integridade e da governança pública, a redução de custos, e a participação social.

Repositórios de Dados Públicos

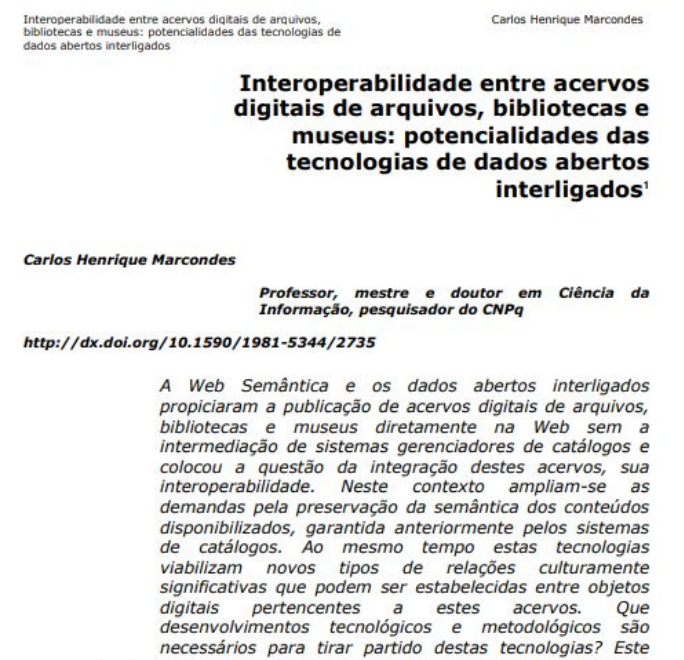
Portal Brasileiro de Dados Abertos e Catálogo Nacional de Dados

Encontre dados publicados pelo governo federal e por governos locais para realizar pesquisas, desenvolver aplicativos e criar novos serviços.

Base dos Dados

Organização não-governamental sem fins lucrativos e open-source que atua para universalizar o acesso a dados de qualidade. Fazemos isso através da criação de ferramentas inovadoras, da produção e difusão do conhecimento e da promoção de uma cultura de transparência e dados abertos.

Leitura para a próxima aula



Interoperabilidade entre acervos digitais de arquivos, bibliotecas e museus: potencialidades das tecnologias de dados abertos interligados[1]

Carlos Henrique Marcondes

Professor, mestre e doutor em Ciência da Informação, pesquisador do CNPq

http://dx.doi.org/10.1590/1981-5344/2735

A Web Semântica e os dados abertos interligados propiciaram a publicação de acervos digitais de arquivos, bibliotecas e museus diretamente na Web sem a intermediação de sistemas gerenciadores de catálogos e colocou a questão da integração destes acervos, sua interoperabilidade. Neste contexto ampliam-se as demandas pela preservação da semântica dos conteúdos disponibilizados, garantida anteriormente pelos sistemas de catálogos. Ao mesmo tempo estas tecnologias viabilizam novos tipos de relações culturamente significativas que podem ser estabelecidas entre objetos digitais pertencentes a estes acervos. Que desenvolvimentos tecnológicos e metodológicos são necessários para tirar partido destas tecnologias? Este

Interoperabilidade entre acervos digitais de arquivos, bibliotecas e museus: potencialidades das tecnologias de dados abertos interligados

Carlos Henrique Marcondes

Disponível em: https://www.scielo.br/j/pci/a/8svGtzqw5HZCrfrPJbRypsb/abstract/?lang=pt

CENTRO DE PESQUISA E FORMAÇÃO
SESC

OBRIGADO!

Até a próxima aula!

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Hora da pausa! Voltamos em:

«20:00-»