

DAS LINGUAGENS DE MARCA ATÉ AS ONTOLOGIAS



NCOR-BR, 2021

O presente material e quaisquer outros recursos que o acompanhem foi produzido pela NCOR-BR para distribuição gratuita e disseminação do tema, sem fins lucrativos.

This work is licensed under



[Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Este trabalho está licenciado sob



[Creative Commons - Atribuição 4.0 Internacional.](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Sumário

| | |
|--|----|
| O que são “marcas” em textos?..... | 4 |
| HTML e XML: aplicações SGML na Web | 7 |
| RDF e RDFS: padrões de metadados na Web..... | 9 |
| OWL: a linguagem para construir ontologias | 12 |

O que são “marcas” em textos?

Os textos escritos e lidos rotineiramente pelas pessoas contêm um tipo de codificação implícita que passa despercebida. Esses códigos envolvem elementos como pontuação, letras maiúsculas e minúsculas, espaços, cabeçalhos, parágrafos, sentenças, dentre outros. Tais elementos ajudam o leitor a identificar as características estruturais necessárias à leitura, por exemplo, determinar o início e o término das palavras ou páginas.

Estamos tão acostumados com isso que em geral nem notamos mais, mas mesmo antes da era digital, **marcas** já eram usadas pelos datilógrafos para entender como o autor desejava o texto. No século XX os textos passaram a ser datilografados e a forma de informar aos datilógrafos sobre os elementos estruturais do texto era usar marcas. As marcas eram padronizadas e transmitiam instruções sobre a apresentação do texto, ou seja, sobre como o autor desejava que fosse escrito (Figura 1).

O termo **marcação** também passou a ser adotado nos anos de 1970 para códigos em textos eletrônicos com função similar às marcas em papel. Assim, “marcar” um documento eletrônico passou a significar a forma de indicar como o conteúdo deve ser interpretado para fins de representação.

Uma **linguagem de marcação** consiste do conjunto de convenções para a codificação de textos eletrônicos. Essa denominação especifica as marcas permitidas, as marcas exigidas, a distinção entre marcas e texto, bem como o significado das marcas.

Figura 1: marcas padronizadas por autores para uso de datilógrafos

| Operational Signs | | Proofreader's Marks* EliteEditing.com | |
|-------------------|--|--|---------------------------------|
| | Delete | | Set in italic type |
| | Close up: delete space | | Set in roman type |
| | Delete and close up (use only when deleting letters within a word) | | Set in boldface type |
| | Let it stand | | Set in lowercase |
| | Insert space | | Set in capital letters |
| | Make space between words equal: | | Set in small capitals |
| | | | Wrong font: set in correct type |

Fonte: Rotner, J. *Proofreaders' Marks*.¹

Enquanto linguagens de programação como C, LISP, Java, etc., fornecem instruções para calcular e executar ações, as linguagens de marcação não executam nada. Foram planejadas para descrever dados objetivando armazenamento, transmissão e processamento. Existem linguagens de marcação com diferentes propósitos, mas a pioneira foi a *Standard Generalized Markup Language* (SGML), um tipo de metalinguagem capaz de gerar outras linguagens.

Figura 1: poema com marcas SGML

```

1 <antologia>
2   <poema>
3     <título> Tenho tanto sentimento </título>
4     <verso>
5       <linha> Tenho tanto sentimento </linha>
6       <linha> Que é frequente persuadi-me </linha>
7       <linha> De que sou sentimental </linha>
8       <linha> Mas reconheço, ao medir-me, </linha>
9     </verso>
10    <verso>
11      <linha> Qual porém é a verdadeira </linha>
12      <linha> E qual errada, ninguém </linha>
13      <linha> Nos saberá explicar </linha>
14      <linha> E vivemos de maneira </linha>
15    </verso>
16  </poema>
17  # mais poemas aqui
18 </antologia>
19
```

O documento marcado é resultado do modelo do SGML, baseado no *elemento*. O elemento é um dispositivo estrutural que delimita ocorrências da unidade textual. A marcação usa os sinais “<” e “>”, respectivamente, como marca inicial e final, delimitando o identificador genérico denominado *nome*. A marca final é identificada pela presença de uma barra antes do nome: “<*nome*> *texto* </*nome*>” (Figura 2).

HTML e XML: aplicações SGML na Web

O SGML é complexo foi substituído por aplicações no escopo da Web:

- *Hypertext Markup Language* (HTML)
- *Extended Markup Language* (XML)

O HTML é uma aplicação originada do SGML para uso na Web, fornecendo um tipo de documento com marcações fixas e cobrindo usos típicos de textos de escritório, como parágrafos, listas, ilustrações, hipertexto e alguns recursos multimídia. O HTML possibilitou que dados fossem apresentados em uma estrutura simples e de fácil leitura, ajudando na disseminação e na popularização da Web.

O XML, por sua vez, foi criado no final dos anos 1990 e tornou-se um “padrão de fato” para intercâmbio de dados digitais. Trata-se de uma versão abreviada do SGML que omite partes mais complexas da linguagem original, permitindo ao autor organizar dados no documento. Foi projetado para carregar metadados e pode conter os dados e a própria descrição da estrutura, ou seja, um *Document Type Definition* (DTD).

O XML não é uma linguagem de marcação com marcas pré-definidas e por isso melhorias no compartilhamento e na recuperação de dados na Web. O XML foi planejado para uso por máquinas, enquanto o HTML foi concebido para gerar páginas para consumo humano. Por causa disso, a apresentação de um texto HTML é de fácil leitura para pessoas, mas não para *sistemas*.

Figura 2– O HTML é para consumo humano
a- marcação HTML em arquivo texto

```
<!DOCTYPE html>
<html>
<head>
<title>Alunos da UFMG</title>
</head>
<body>

<h1>Lista de alunos da pós graduação</h1>
<p><b>João</b>, 30 anos, <i>joao@ufmg.br</i></p>
<p><b>Cláudia</b>, 25 anos, <i>kk@ufmg.br</i></p>
<p><b>José</b>, 30 anos, <i>ze@ufmg.br</i></p>

</body>
</html>
```

b- resultado lido no navegador

Alunos da pós graduação

João, 30 anos, *joao@ufmg.br*

Cláudia, 25 anos, *kk@ufmg.br*

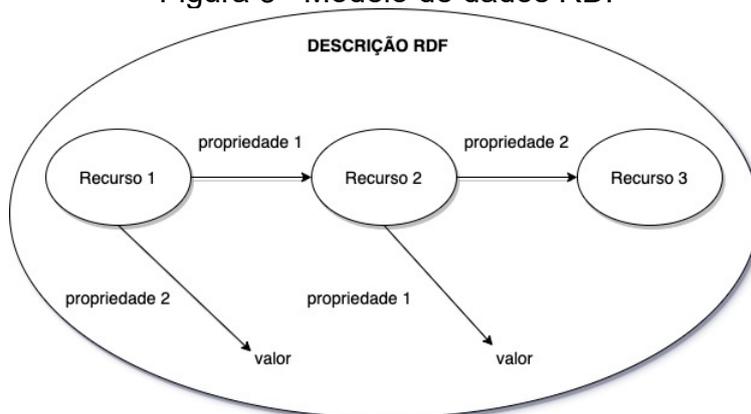
José, 30 anos, *ze@ufmg.br*

RDF e RDFS: padrões de metadados na Web

As limitações do XML levaram aos padrões de metadados – o *Resource Description Framework (RDF)*² e o *Resource Description Framework Schema (RDFS)*³ – capazes de intercambiar e reutilizar dados e metadados. Esses padrões usam a sintaxe XML para impor limitações adicionais aos dados.

O RDF se mostrou um meio eficaz para a publicação de vocabulários legíveis por pessoas e por computadores, além de permitir a distribuição de metadados em diferentes comunidades. O poder do RDF reside no mecanismo do *namespace* proveniente do XML que permitiu o compartilhamento de vocabulários a respeito de diversos assuntos pela reutilização de *Uniform Resource Identifier (URI)*.

Figura 3– Modelo de dados RDF



O RDF possui um modelo de dados simples, mas eficiente, baseado em *recursos*. (Figura 3). Um recurso consiste de objetos identificados por uma URI. Nas primeiras versões, as propriedades RDF eram chamadas *propriedades-tipo* e relacionavam valores a recursos. Uma propriedade também pode ser identificada por uma URI e ser unária, por exemplo o “peso” em “peso da pessoa”, ou binária, por exemplo “escrito por” em “livro escrito-por Jorge”.

Os valores em RDF podem assumir duas formas i) *strings*, ou seja, qualquer conjunto de caracteres; ou ii) outros recursos, os quais podem ter suas próprias propriedades. Talvez a noção mais importante para se entender inicialmente na sintaxe RDF é a *descrição* RDF. A coleção dos recursos, propriedades e valores é que forma a *descrição* RDF (Figura 3).

Um exemplo da marcação RDF exibe uma marcação um pouco mais complexa. Observe que a URI agora vai para dentro da marca na forma de um namespace, o qual a abrevia (Figura 4).

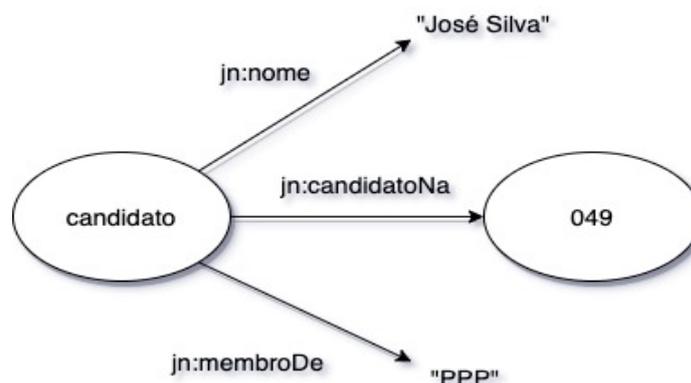
Figura 4– Marcação RDF para descrever uma eleição
a- código RDF

```

1 <rdf:Description
2   xmlns:rdf="http://www.w3c.org/1999/02/22-rdf-syntax-ns#"
3   xmlns:jn="http://www.jornal.com.br/candidato">
4   xmlns:tre="http://www.eleitoral.org.br/elege/">
5     rdf:about="http://jornal.com.br/candidato/">
6       <jn:nome> José Silva <dc:nome>
7       <jn:CandidatoNa
8         rdf:resource="http://www.eleitoral.org.br/elege/zona/049">
9     </jn:CandidatoNa>
10    <jn:membroDe>
11      PPP
12    </jn:membroDe>
13 </rdf:Description>

```

b- representação gráfica



O RDFS, por sua vez, assegura que a descrição RDF use os significados adequados porque introduz *restrições*. Para fazer isso, a linguagem verifica: se o objeto é adequado ao sujeito e se o sujeito é adequado ao predicado.

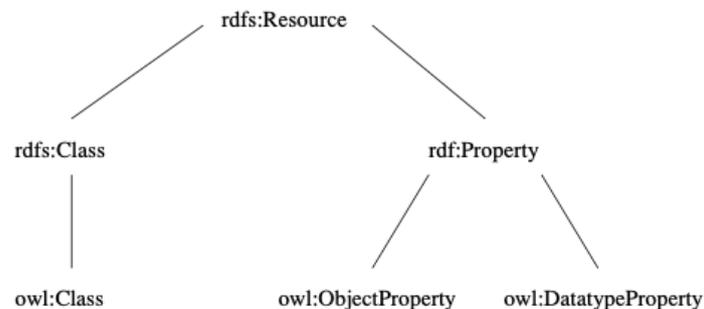
Para resolver o primeiro problema é preciso indicar o tipo do sujeito, para evitar, por exemplo, que uma data de aniversário seja atribuída a uma pedra. Para resolver o segundo problema, o RDFS é capaz de indicar que tipo de sujeito pode ser associado aos predicados, por exemplo, apenas livros (além de documentos correlatos como revistas científicas, etc) podem ter a propriedade ISBN.

OWL: a linguagem para construir ontologias

As iniciativas de marcação continuaram na direção de uma linguagem para ontologias, o que resultou na *Web Ontology Language (OWL)*⁴. De fato, a Web, e mesmo sistemas locais, exibem situações em que existe demanda por expressividade além daquela que RDF e RDFS oferecem. A OWL representa mais uma evolução na direção de prover significado e se mantêm a melhor solução para a Web Semântica.

A OWL faz uso da RDFS em larga escala, pois usa o RDF como sintaxe. Elementos importantes do OWL são especializações de elementos RDF (Figura 5).

Figura 5– Relacionamento entre classes OWL e RDF/RDFS



O OWL possui novos elementos incorporados como elementos para classes, instâncias e para propriedades; as restrições sobre propriedades e sobre cardinalidade.

A OWL tem diversas sintaxes e pelos menos as listadas a seguir são muito disseminadas. As vantagens, desvantagens e usos é assunto para material posterior.

- Sintaxe *RDF/XML*
- Sintaxe *Terse RDF Triple Language (Turtle)*
- Sintaxe *OWL/XML*
- Sintaxe *OWL Functional*
- Sintaxe *Manchester OWL*

ALMEIDA, M.B. Ontologia em Ciência da Informação: Tecnologia e Aplicações. Curitiba: CRV, 2021, 310 p. Coleção Representação do Conhecimento em Ciência da Informação: Volume 2.

Esse material é gratuito, mas você pode ajudar na divulgação citando a fonte:

NCOR-BR. (2021). Das linguagens de marca até as ontologias. Caderno 9. Disponível na internet em <<http://ncor-br.org>>

¹ Disponível na internet em: <<https://medium.com/@jenniferotner/proofreaders-marks-965ba7e9a48c>>. Acesso em 11/01/2021

² Disponível na Internet em: <<https://www.w3.org/RDF/>>. Acesso em 15 de março de 2021.

³ Disponível na Internet em: <<https://www.w3.org/TR/rdf-schema/>>. Acesso em 15 de março de 2021.

⁴ Disponível na internet em: <<https://www.w3.org/OWL/>>. Acesso em 15/03/2021.