



Universidade Estadual Paulista (UNESP)
Faculdade de Filosofia e Ciências, Campus de Marília
Programa de Pós-Graduação em Ciência da Informação

Fabício Silva Assumpção

Modelo para a publicação de dados de autoridade como *Linked Data*

Marília
2018

Fabício Silva Assumpção

Modelo para a publicação de dados de autoridade como *Linked Data*

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da UNESP, Campus de Marília, como requisito parcial para a obtenção do título de doutor em Ciência da Informação.

Linha de Pesquisa: Informação e Tecnologia.

Orientadora: Plácida Leopoldina Ventura Amorim da Costa Santos

Marília
2018

fabricioassumpcao.com
assumpcao.f@gmail.com

A851m Assumpção, Fabrício Silva.
Modelo para a publicação de dados de autoridade como Linked Data /
Fabrício Silva Assumpção. – Marília, 2018.
208 f. ; 30 cm.

Orientador: Plácida Leopoldina Ventura Amorim da Costa Santos.
Tese (Doutorado em Ciência da Informação) – Universidade Estadual
Paulista (Unesp), Faculdade de Filosofia e Ciências, 2018.
Bibliografia: f. 193-208.

1. Controle de autoridade. 2. Dados ligados. 3. Web semântica. 4.
Catalogação descritiva. 5. Tecnologia da informação. I. Título.

CDD 005.73

Ficha catalográfica elaborada por
André Sávio Craveiro Bueno
Bibliotecário CRB 8/8211

Fabrcio Silva Assumpção

Modelo para a publicação de dados de autoridade como *Linked Data*

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da UNESP, Campus de Marília, como requisito parcial para a obtenção do título de doutor em Ciência da Informação.

Banca examinadora:

Prof.^a Dr.^a Plácida Leopoldina Ventura Amorim da Costa Santos
Universidade Estadual Paulista (UNESP) – Orientadora

Dr.^a Ana Maria Pereira
Universidade do Estado de Santa Catarina (UDESC)

Dr. José Eduardo Santarem Segundo
Universidade de São Paulo (USP)

Prof. Dr. Ricardo Cesar Gonçalves Sant'Ana
Universidade Estadual Paulista (UNESP)

Dr.^a Zaira Regina Zafalon
Universidade Federal de São Carlos (UFSCar)

UNESP, Campus de Marília, 05 de fevereiro de 2018.

Agradecimentos

Agradeço primeiramente à minha mãe. Agradeço também:

À professora Plácida, pelas inúmeras contribuições à minha formação no decorrer da graduação, do mestrado e do doutorado e pelas oportunidades e liberdades a mim proporcionadas para concretizar meus projetos de pesquisa.

Aos professores membros da banca, Ana Maria Pereira, José Eduardo Santarem Segundo, Ricardo Cesar Gonçalves Sant'Ana e Zaira Regina Zafalon, e suplentes, Rachel Cristina Vesu Alves, Rogério Aparecido Sá Ramalho e Silvana Aparecida Borsetti Gregorio Vidotti, por sua disposição em participar da avaliação deste trabalho e por suas contribuições para minha formação.

Aos diversos colegas, alunos e professores, do Grupo de Pesquisa Novas Tecnologias em Informação (GPNTI), especialmente aos mais próximos com os quais convivi por tanto tempo no Laboratório de Tecnologias Informacionais.

Aos professores do Departamento de Ciência da Informação e do Programa de Pós-Graduação em Ciência da Informação da UNESP, com quem muito aprendi ao longo dos anos. Ao professor Walter Moreira, também pela oportunidade de realizar o estágio docência na disciplina “Linguagens documentais alfabéticas”.

Aos colegas de trabalho do Repositório Institucional UNESP e da Biblioteca do Setor Litoral da Universidade Federal do Paraná (UFPR), por seu companheirismo e pelas oportunidades de crescimento pessoal e profissional que me proporcionaram.

À UNESP, por tudo que me proporcionou durante a graduação, o mestrado e o doutorado, e à CAPES, pela bolsa concedida entre março e outubro de 2016.

A tantos outros que, direta ou indiretamente, estiveram comigo durante os 10 (dez!) anos dessa caminhada quase ininterrupta da graduação ao doutorado.

E ao Jaider, por tudo!

Instant electronic access to digital information is the single most distinguishing attribute of the information age. The elaborate retrieval mechanisms that support such access are a product of technology. But technology is not enough. The effectiveness of a system for accessing information is a direct function of the intelligence put into organizing it. Just as the practical field of engineering has theoretical physics as its underlying base, the design of systems for organizing information rests on an intellectual foundation.

— ELAINE SVENONIUS (2000, p. ix)

Resumo

A Ciência da Informação tem entre suas preocupações o acesso à informação e aos recursos informacionais, sendo, portanto, objetos de seu interesse os instrumentos utilizados para esse acesso, tais como os catálogos, que incluem dados bibliográficos (representações de recursos informacionais) e dados de autoridade (representações das entidades associadas aos recursos informacionais, tais como pessoas, entidades coletivas e conceitos). A proposta de criação de uma Web Semântica, em que os dados sejam processados não somente por sua sintaxe, mas também por sua semântica, tem impulsionado o desenvolvimento de um conjunto de tecnologias para a representação de dados na Web, assim como para a consulta a esses dados e o raciocínio computadorizado a partir deles. O uso de algumas dessas tecnologias para a publicação e o relacionamento de dados levou ao surgimento do conceito de *Linked Data*, e o anseio por sua aplicação na Ciência da Informação deu origem a projetos para a publicação de dados de autoridade como *Linked Data*. No entanto, observa-se que esses projetos, ainda em estágios iniciais ou experimentais, carecem de um quadro teórico construído na Ciência da Informação que possa orientar quanto às políticas, aos procedimentos e às tecnologias empregadas na publicação desses dados. Assim, partindo do problema “como publicar dados de autoridade como *Linked Data*?” são delineados o objetivo geral – propor, a partir do conceito das funções dos dados de autoridade nos catálogos e de seus benefícios nos ambientes de *Linked Data*, um modelo de publicação de dados de autoridade como *Linked Data* compreendendo políticas, procedimentos e tecnologias – a tese e a hipótese desta pesquisa, de caráter bibliográfico, documental e metodológico. Para o alcance desse objetivo, primeiramente é conduzida uma revisão de literatura acerca do controle de autoridade e do desenvolvimento dos dados de autoridade nos catálogos em livros, em fichas e digitais, com destaque para os modelos conceituais FRAD e FRSAD que sintetizam as funções desempenhadas pelos dados de autoridade. Em seguida, são apresentados o conceito de *Linked Data* e as principais tecnologias da Web Semântica relacionadas a ele: URIs, RDF, RDFS e OWL, apresentação essa que serve de base para a descrição de três vocabulários que podem ser utilizados na publicação de dados de autoridade (SKOS, MADS/RDF e *RDA Element Sets*), de iniciativas para a publicação desses dados (*LC Linked Data Service*, *datos.bne.es*, *data.bnf.fr*, VIAF e AGROVOC) e dos potenciais benefícios da publicação dos dados de autoridade como *Linked Data*. A partir dos resultados dessa revisão de literatura, é proposto o modelo para a publicação de dados de autoridade como *Linked Data*, compreendendo as etapas de planejamento; modelagem e mapeamento; tratamento, relacionamento e conversão; publicação; e *feedback* e retroalimentação. Após a descrição de cada etapa do modelo, com suas políticas, procedimentos e tecnologias, são apresentadas considerações finais sobre os resultados alcançados e sobre o modelo proposto.

Palavras-chave: Dados de autoridade. Controle de autoridade. *Linked Data*. Web Semântica. Catálogos. Catalogação descritiva. Informação e tecnologia.

Abstract

One of the issues that Information Science is concerned with is the access to information resources, therefore the studies in this area include the library catalogs, which comprise bibliographic data (representations of information resources) and authority data (representations of the entities related to the information resources, such as persons, corporate body and concepts). The proposal of a Semantic Web in which data are processed not just by their syntax but also by their semantics has led to the development of a set of technologies for publishing and linking data on the Web, as well as technologies for data querying and for the computerized reasoning. The use of a subset of these technologies for publishing and linking data has led to the Linked Data concept. The experiments with these technologies in Information Science have given rise to projects for publishing authority data as Linked Data. However, we observed that these projects, still in initial or experimental stages, are not based on a theoretical framework developed in Information Science that can guide them regarding to the policies, to the procedures and to the technologies used in the publication of these data. Thus, starting from the question “how to publish authority data as Linked Data?” we define the main goal – conceptualize the functions of authority data in the catalogs and their benefits in Linked Data environments in order to propose a model for authority data publishing, as well as highlighting its policies, procedures and technologies – the thesis and the hypothesis of this research. To achieve this goal, first we made a literature review about authority control and the development of authority data in books, cards and digital catalogs, remarking the FRAD and FRSAD conceptual models which synthesize the functions of authority data. Then, we present the Linked Data concept and the main Semantic Web technologies related to it: URIs, RDF, RDFS and OWL; starting from this introduction, we describe three vocabularies that can be used to publishing authority data (SKOS, MADS/RDF and RDA Element Sets), some initiatives (Library of Congress Linked Data Service, datos.bne.es, data.bnf.fr, VIAF and AGROVOC) and the advantages of authority data published as Linked Data. Based on the results of this literature review, we propose a model for publishing authority data as Linked Data, The model comprises the following stages: planning; modeling and mapping; processing, linking and conversion; publishing; and feedback. After describing each of these stages, with its policies, procedures and technologies, we present the conclusions about the results and about the proposed model.

Keywords: Authority data. Authority control. Linked Data. Semantic Web. Catalogs. Descriptive cataloging. Information and technology.

Lista de figuras

Figura 1 – Relações entre as entidades e seus nomes	26
Figura 2 – Relações entre entidades, nomes e recursos	27
Figura 3 – Sistema de autoridade com seus componentes e relacionamentos.....	33
Figura 4 – Ficha de autoridade para um nome pessoal	46
Figura 5 – Modelo de dados dos catálogos digitais.....	54
Figura 6 – Base fundamental do FRAD	62
Figura 7 – Relacionamentos entre as entidades do FRAD	65
Figura 8 – Relacionamento entre as entidades obra, <i>thema</i> e <i>nomen</i>	68
Figura 9 – Tecnologias da Web Semântica.....	74
Figura 10 – <i>Linking Open Data cloud diagram</i>	76
Figura 11 – URIs identificando objetos do mundo real e documentos	81
Figura 12 – Grafo RDF	84
Figura 13 – Formatos para a serialização do RDF.....	85
Figura 14 – Serialização do RDF utilizando o formato <i>Turtle</i>	87
Figura 15 – Declarações utilizando a propriedade <i>rdf:type</i> serializadas em <i>Turtle</i>	95
Figura 16 – Declarações utilizando propriedades e classes do <i>RDF Schema</i>	96
Figura 17 – Declarações em RDF utilizando algumas propriedades da OWL	100
Figura 18 – Categorias de dados das bibliotecas.....	104
Figura 19 – Exemplos de classes e propriedades do SKOS.....	108
Figura 20 – Exemplos de classes e propriedades do MADS/RDF	113
Figura 21 – Exemplos de propriedades dos <i>RDA Element Sets</i>	118
Figura 22 – Modelagem utilizada no portal <i>datos.bne.es</i>	126
Figura 23 – Arquitetura global do projeto <i>data.bnf.fr</i>	129
Figura 24 – <i>UNLV Linked Data Project</i>	145
Figura 25 – Modelo para a publicação de dados de autoridade como <i>Linked Data</i>	147
Figura 26 – Machado de Assis na <i>datos.bne.es</i>	159
Figura 27 – Modelo de dados para dados de autoridade	161
Figura 28 – Uso do código 303 com negociação de conteúdo	168
Figura 29 – Descrição em RDF de um conjunto de dados	180
Figura 30 – Página sobre o ano de 1973.....	184
Figura 31 – Página sobre a cidade de Roma.....	185

Lista de quadros

Quadro 1 – Entidades e atributos definidos no FRAD	63
Quadro 2 – Entidades e atributos definidos no FRSAD	67
Quadro 3 – Exemplos de vocabulários utilizados em declarações RDF	92
Quadro 4 – Iniciativas para a publicação de dados de autoridade como <i>Linked Data</i>	134
Quadro 5 – Metas e ações do projeto <i>datos.bne.es</i>	143
Quadro 6 – Visão geral das fases do <i>UNLV Linked Data Project</i>	144
Quadro 7 – Melhores práticas para os dados na Web	146
Quadro 8 – Classificação das licenças a partir de suas restrições.....	151
Quadro 9 – Licenças <i>Creative Commons</i>	152
Quadro 10 – Exemplo de mapeamento	164

Lista de siglas

AACR	<i>Anglo-American Cataloguing Rules</i>
AACR2r	<i>Anglo-American Cataloguing Rules, 2nd edition, 2002 revision</i>
AFS	<i>American Folklore Society</i>
AGROVOC	<i>Agriculture with Vocabulary</i>
ALA	<i>American Library Association</i>
ANSI	<i>American National Standards Institute</i>
API	<i>Application Programming Interface</i>
ARK	<i>Archival Resource Key</i>
BIBFRAME	<i>Bibliographic Framework</i>
BNE	<i>Biblioteca Nacional de España</i>
BnF	<i>Bibliothèque nationale de France</i>
CAPES	<i>Coordenação de Aperfeiçoamento de Pessoal de Nível Superior</i>
CC-0	<i>Creative Commons – Public Domain Dedication</i>
CC-BY	<i>Creative Commons – Attribution</i>
CD-ROM	<i>Compact Disc Read-Only Memory</i>
CSV	<i>Comma-Separated Values</i>
DCMI	<i>Dublin Core Metadata Initiative</i>
DNB	<i>Deutsche Nationalbibliothek</i>
DOI	<i>Digital Object Identifier</i>
EUA	<i>Estados Unidos da América</i>
FAO	<i>Food and Agriculture Organization</i>
FOAF	<i>Friend of a Friend</i>
FRAD	<i>Functional Requirements for Authority Data</i>
FRANAR	<i>Working Group on Functional Requirements and Numbering of Authority Records</i>
FRBR	<i>Functional Requirements for Bibliographic Records</i>
FRSAD	<i>Functional Requirements for Subject Authority Data</i>
FRSAR	<i>Working Group on the Functional Requirements for Subject Authority Records</i>
FTP	<i>File Transfer Protocol</i>
GEMET	<i>General Multilingual Environmental Thesaurus</i>
HTML	<i>Hypertext Markup Language</i>
HTTP	<i>Hypertext Transfer Protocol</i>

IFLA	<i>International Federation of Library Associations and Institutions</i>
ISADN	<i>International Standard Authority Data Number</i>
ISBN	<i>International Standard Book Number</i>
ISNI	<i>International Standard Name Identifier</i>
ISO	<i>International Organization for Standardization</i>
ISSN	<i>International Standard Serial Number</i>
JSON	<i>JavaScript Object Notation</i>
JSON-LD	<i>JavaScript Object Notation for Linked Data</i>
LC	<i>Library of Congress</i>
LCSH	<i>Library of Congress Subject Headings</i>
LLD	<i>Library Linked Data</i>
LOD	<i>Linked Open Data</i>
MADS	<i>Metadata Authority Description Schema</i>
MADS/RDF	<i>Metadata Authority Description Schema in RDF</i>
MARC	<i>MAchine-Readable Cataloging</i>
MODS	<i>Metadata Object Description Schema</i>
N3	<i>Notation 3</i>
NISO	<i>National Information Standards Organization</i>
NUNA	<i>Non-Unique Name Assumption</i>
OCLC	<i>Online Computer Library Center</i>
ONIX	<i>Online Information Exchange</i>
OPAC	<i>Online Public Access Catalog</i>
OWL	<i>Web Ontology Language</i>
PHP	<i>PHP Hypertext Preprocessor</i>
PPGCI	<i>Programa de Pós-Graduação em Ciência da Informação</i>
PREMIS	<i>Preservation Metadata Maintenance Activity</i>
RAMEAU	<i>Répertoire d'autorité-matière encyclopédique et alphabétique unifié</i>
RDA	<i>Resource Description and Access</i>
RDF	<i>Resource Description Framework</i>
RDFS	<i>RDF Schema</i>
SKOS	<i>Simple Knowledge Organization System</i>
SKOS-XL	<i>SKOS eXtension for Labels</i>
SPARQL	<i>SPARQL Protocol and RDF Query Language</i>
SUDOC	<i>Système Universitaire de Documentation</i>
SWAD-Europe	<i>Semantic Web Advanced Development for Europe</i>

UNESP	Universidade Estadual Paulista “Júlio de Mesquita Filho”
UNIMARC	<i>Universal Machine Readable Cataloging</i>
URI	<i>Uniform Resource Identifier</i>
URL	<i>Uniform Resource Locators</i>
VIAF	<i>Virtual International Authority File</i>
VoID	<i>Vocabulary of Interlinked Datasets</i>
W3C	<i>World Wide Web Consortium</i>
XML	<i>Extensible Markup Language</i>
XSLT	<i>Extensible Stylesheet Language for Transformation</i>

Sumário

1	Introdução.....	14
1.1	Problema, tese e hipótese	15
1.2	Objetivos.....	16
1.3	Justificativa, relevância e motivação da pesquisa	17
1.4	Metodologia	18
1.5	Estrutura da pesquisa	19
2	Catálogos e controle de autoridade.....	21
2.1	Controle de autoridade.....	24
2.2	Controle de autoridade nos catálogos em livros e em fichas.....	35
2.3	Controle de autoridade nos catálogos digitais	48
2.4	<i>Functional Requirements for Authority Data (FRAD) e Functional Requirements for Subject Authority Data (FRSAD)</i>	<i>59</i>
3	Web Semântica, suas tecnologias e <i>Linked Data</i>.....	71
3.1	<i>Uniform Resource Identifier (URI) Hypertext Transfer Protocol (HTTP).....</i>	<i>78</i>
3.2	<i>Resource Description Framework (RDF).....</i>	<i>82</i>
3.3	<i>RDF Schema (RDFS) e Web Ontology Language (OWL)</i>	<i>91</i>
4	<i>Linked Data</i> no domínio bibliográfico: os dados de autoridade em foco....	103
4.1	Vocabulários para a descrição de dados de autoridade.....	106
4.1.1	<i>Simple Knowledge Organization System (SKOS).....</i>	<i>107</i>
4.1.2	<i>Metadata Authority Description Schema in RDF (MADS/RDF).....</i>	<i>111</i>
4.1.3	<i>RDA Element Sets.....</i>	<i>115</i>
4.2	Dados de autoridade publicados como <i>Linked Data</i>	121
4.2.1	<i>Library of Congress Linked Data Service.....</i>	<i>121</i>
4.2.2	<i>datos.bne.es.....</i>	<i>123</i>
4.2.3	<i>data.bnf.fr.....</i>	<i>127</i>
4.2.4	<i>Virtual International Authority File (VIAF)</i>	<i>130</i>
4.2.5	<i>AGROVOC Multilingual agricultural thesaurus</i>	<i>131</i>
4.3	Benefícios dos dados de autoridade publicados como <i>Linked Data</i>	136

5	Modelo para a publicação de dados de autoridade como <i>Linked Data</i>.....	142
5.1	Planejamento.....	147
5.1.1	Seleção e análise dos conjuntos de dados	148
5.1.2	Escolha da licença.....	149
5.1.3	Escolha dos conjuntos de dados para relacionamento	155
5.2	Modelagem e mapeamento dos dados.....	157
5.2.1	Modelagem dos dados.....	157
5.2.2	Mapeamento	162
5.2.3	Descrição do vocabulário utilizado.....	166
5.2.4	Definição de padrões para os URIs	167
5.3	Tratamento, relacionamento e conversão dos dados.....	170
5.3.1	Tratamento dos dados	170
5.3.2	Relacionamento com outros conjuntos de dados	175
5.3.3	Conversão para RDF	177
5.4	Publicação dos dados	179
5.4.1	Descrição do conjunto de dados	179
5.4.2	Dados acessíveis aos softwares.....	180
5.4.3	Dados acessíveis aos humanos.....	183
5.5	<i>Feedback</i> e retroalimentação do arquivo de autoridade	185
6	Considerações finais.....	187
	Referências	193

1 Introdução

Esta tese é resultado da pesquisa desenvolvida no Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), Campus de Marília, na linha de pesquisa “Informação e Tecnologia” e contou, durante um período, com o financiamento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por meio da concessão de bolsa.

A Ciência da Informação, no entendimento de Borko (1968, p. 3, tradução nossa), preocupa-se “com o corpo de conhecimento relacionado à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação”. Para cobrir esses diversos aspectos e atividades relacionadas à informação, a Ciência da Informação tem traçado relações com outras áreas, entre elas a Biblioteconomia e a Ciência da Computação (SARACEVIC, 1995).

Para Saracevic (1995, p. 4), a Ciência da Informação e a Biblioteconomia têm uma forte base em comum em razão de compartilharem seu papel social e sua preocupação com o uso efetivo de registros ou recursos informacionais. De forma semelhante, a Ciência da Informação está inexoravelmente conectada à tecnologia da informação (SARACEVIC, 1995, p. 2).

Nota-se também a relação que a Ciência da Informação mantém com a Catalogação que, originada da prática biblioteconômica de criação e de gestão de catálogos, hoje se ocupa com um conjunto de teorias, de instrumentos e de tecnologias voltados à representação de recursos informacionais. Entre os objetos de preocupação da Catalogação e, de forma geral, da Ciência da Informação, estão os catálogos bibliográficos e seus componentes – sistemas de informação que têm evoluído desde listagens criadas em ambientes analógicos até complexos bancos de dados digitais que almejam a integração com outros serviços de informação também disponíveis e acessíveis na Web. É nesse sentido que se concorda com Santos e Sant’Ana (2013, p. 200):

A ciência da informação refere-se à atividade direcionada à pesquisa de princípios e métodos que são partes da análise, do projeto e da evolução dos sistemas de informação. Nesses sistemas, os elementos constituintes são o ambiente, as pessoas, os recursos informacionais, as tecnologias e os procedimentos.

Enquanto sistemas de informação, os catálogos, de modo geral, proveem representações de recursos informacionais com o propósito de permitir aos seus usuários a localização, a identificação, a seleção e o acesso a esses recursos, assim como a navegação pelo catálogo e para além dele (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009). Durante sua evolução, nota-se nos catálogos a presença de dados que, ainda que não se constituam como representações de recursos informacionais, apoiam o uso e a gestão dos catálogos. Esses dados, definidos inicialmente e para os propósitos desta introdução como dados referentes às entidades relacionadas aos recursos informacionais representados em um determinado catálogo, são denominados dados de autoridade.

O surgimento da Web na década de 1990 trouxe novos meios para a publicação, o acesso e o intercâmbio de dados em ambientes digitais. A partir do início do século XXI, a proposta de criação de uma Web em que os dados sejam processados não somente por sua sintaxe, mas também por sua semântica – a chamada Web Semântica (BERNERS-LEE; HENDLER; LASSILA, 2001) – tem conduzido ao desenvolvimento de um conjunto de tecnologias para a representação dos dados na Web, assim como para a consulta a esses dados e o raciocínio computadorizado a partir deles. A utilização de um subconjunto dessas tecnologias para a publicação e o relacionamento de dados na Web levou ao surgimento do conceito de *Linked Data* (BERNERS-LEE, 2006).

O anseio pela aplicação de tecnologias da Web Semântica no âmbito da Ciência da Informação leva instituições a iniciarem projetos para a publicação de dados de autoridade como *Linked Data*. No entanto, observa-se que esses projetos, ainda em estágios iniciais ou experimentais, não dispõem de um quadro teórico construído na Ciência da Informação que possa orientar quanto às políticas, aos procedimentos e às tecnologias empregadas na publicação desses dados. A partir da observação dessa necessidade teórica e prática, são delineados o problema, a tese, a hipótese e os objetivos desta pesquisa.

1.1 Problema, tese e hipótese

Esta pesquisa inicia-se a partir do **problema**: como publicar dados de autoridade como *Linked Data*? Esse problema de pesquisa é mais bem explicitado nos questionamentos:

- Quais funções os dados de autoridade desempenham nos catálogos bibliográficos?
- Quais políticas e práticas têm sido adotadas nas iniciativas para a publicação de dados de autoridade como *Linked Data*?
- Quais padrões de metadados (vocabulários de classes e de propriedades) podem ser utilizados na publicação de dados de autoridade como *Linked Data*?
- Quais benefícios a publicação de dados de autoridade como *Linked Data* pode trazer para os catálogos bibliográficos e para a Web Semântica?
- Como estão sendo conduzidas as iniciativas para a publicação de dados de autoridade como *Linked Data*?

Diante desse problema de pesquisa, a **tese** defendida é que: o conhecimento acerca das funções dos dados de autoridade no contexto dos catálogos bibliográficos, aliado ao conhecimento das tecnologias da Web Semântica, possibilita a publicação dos dados de autoridade como *Linked Data* de modo que tais dados possam desempenhar, em nível macro na Web Semântica, as funções que desempenham em nível micro nos catálogos bibliográficos.

Desse modo, a **hipótese** levantada é que um modelo, construído a partir (1) do entendimento das funções dos dados de autoridade nos catálogos bibliográficos, (2) do entendimento das tecnologias da Web Semântica e (3) da análise das principais iniciativas e vocabulários para a publicação de dados de autoridade como *Linked Data*, auxiliará na publicação de dados de autoridade como *Linked Data*.

Para verificar essa hipótese, são elencados um objetivo geral e quatro objetivos específicos, descritos no item seguinte.

1.2 Objetivos

O **objetivo geral** é propor, a partir do conceito das funções dos dados de autoridade nos catálogos e de seus benefícios nos ambientes de *Linked Data*, um modelo de publicação de dados de autoridade como *Linked Data* compreendendo políticas, procedimentos e tecnologias.

Os **objetivos específicos** são:

1. Conceituar as funções dos dados de autoridade nos catálogos bibliográficos.

2. Contextualizar as principais tecnologias da Web Semântica em relação à publicação de *Linked Data*.
3. Identificar os potenciais benefícios da publicação de dados de autoridade como *Linked Data* para os catálogos bibliográficos e para a Web.
4. Identificar as principais iniciativas para a publicação de dados de autoridade como *Linked Data* e descrever os principais vocabulários que podem ser utilizados na publicação desses dados.

1.3 Justificativa, relevância e motivação da pesquisa

As tecnologias da Web Semântica e, mais recentemente, o emprego dessas tecnologias na publicação de *Linked Data*, têm sido objetos de estudo da Ciência da Informação em âmbito nacional e internacional. Na Catalogação, essas tecnologias e o conceito de *Linked Data* têm figurado nos últimos anos em decorrência da criação de vocabulários de classes e de propriedades a partir de padrões de metadados já existentes, com o intuito de habilitar os dados bibliográficos e de autoridade para usos, na Web, além do domínio bibliográfico e de suas ferramentas. Assim, considerando as possíveis contribuições de um modelo para a publicação de dados de autoridade como *Linked Data*, esta pesquisa justifica-se por sua relevância científica, acadêmica, profissional e social.

Esta pesquisa justifica-se por sua relevância científica por conceituar as funções dos dados de autoridade nos catálogos bibliográficos e os potenciais benefícios da publicação desses dados como *Linked Data*, contribuindo, desse modo, para um amadurecimento da compreensão dos conceitos de controle de autoridade, dados de autoridade, *Linked Data* e Web Semântica no âmbito da Ciência da Informação. Na esfera acadêmica, esta pesquisa e os produtos dela derivados passam a integrar o corpo de publicações nacionais da Ciência da Informação, contribuindo para levantamentos acerca do estado da arte dos temas aqui estudados e, sobretudo, para o ensino de Catalogação e das tecnologias da Web Semântica.

Esta pesquisa justifica-se também por sua relevância profissional por abordar possibilidades de aplicação das tecnologias da Web Semântica para a representação e o acesso aos dados bibliográficos e de autoridade, podendo, portanto, contribuir para a criação, a gestão e o aperfeiçoamento dos catálogos bibliográficos de modo a melhorar

sua eficiência enquanto ferramentas de acesso à informação. De forma específica, esta pesquisa auxiliará no posicionamento e nas ações das instituições gestoras de catálogos bibliográficos frente às iniciativas de *Linked Data*. Entende-se aqui que essas contribuições, de forma direta ou indireta, impactam na qualidade dos serviços e das ferramentas ofertadas aos usuários dos sistemas de informação, em especial dos catálogos, portanto, estendem-se também à esfera social.

A originalidade desta pesquisa reside na conceituação acerca das funções dos dados de autoridade, nos catálogos bibliográficos e no *Linked Data*, que culmina em um produto inédito (o modelo para a publicação de dados de autoridade como *Linked Data*).

Entre os fatores que motivaram a proposição e o desenvolvimento desta pesquisa estão:

1. o estudo inicial acerca do controle de autoridade, realizado no trabalho de conclusão do curso de graduação;
2. a investigação realizada na dissertação de mestrado, que culminou na proposição de um modelo para a conversão de registros em XML para os Formatos MARC 21 utilizando folhas de estilo XSLT;
3. a observação de iniciativas internacionais para a publicação de *Linked Data* no âmbito da Catalogação;
4. a ausência de experimentos com as tecnologias da Web Semântica e as práticas de *Linked Data* envolvendo bibliotecas brasileiras; e
5. o constante questionamento: “como publicar os dados da minha biblioteca como *Linked Data*?”.

1.4 Metodologia

Esta pesquisa, de natureza qualitativa, é considerada bibliográfica, documental e metodológica. É bibliográfica por recorrer à literatura da Ciência da Informação e da Ciência da Computação para a discussão de temas como os catálogos, os dados de autoridade, a Web Semântica e o *Linked Data*. É também documental, por recorrer a fontes documentais, em sua maioria documentos técnicos, especificando tecnologias e descrevendo iniciativas. Por fim, é também uma pesquisa metodológica, por ter em seu objetivo a proposição de um modelo para a publicação de dados de autoridade como *Linked Data*. A pesquisa metodológica, segundo Demo (1995, p. 13), é aquela “dedicada a

indagar por instrumentos, por caminhos, por modos de se fazer ciência, ou a produzir técnicas de tratamento da realidade, ou a discutir abordagens teórico-práticas”.

Para os procedimentos bibliográficos fez-se uso de bases de dados nacionais e internacionais acerca dos temas de investigação, cobrindo publicações nos idiomas português, inglês e espanhol, sem restrições cronológicas ou tipológicas. Para os procedimentos documentais, que subsidiaram principalmente o Capítulo 3 e as Seções 4.1 e 4.2, priorizou-se o uso de documentos emitidos pelas instituições responsáveis pelas tecnologias e iniciativas estudadas.

Nas revisões de literatura, em especial naquela conduzida no Capítulo 2, a ênfase nos dados de autoridade que representam principalmente entidades do tipo pessoa e entidade coletiva, em detrimento daqueles que representam conceitos (“termos tópicos”), ocorre de forma consciente, partindo (1) da necessidade de delimitação do objeto de estudo, com vistas a garantir a exequibilidade da pesquisa, e (2) dos referenciais teóricos adotados.

A escolha pelas tecnologias descritas no Capítulo 3 e na Seção 4.1 levou em conta principalmente a aplicabilidade dessas tecnologias na publicação de dados de autoridade como *Linked Data*, assim, tecnologias da Web Semântica relacionadas às camadas de lógica, de regras, de verdade, de segurança e de prova, por exemplo, não são abordadas, o mesmo ocorre com vocabulários que, mesmo tendo sido criados para a descrição de pessoas e de entidades coletivas, não têm em seu escopo a pretensão de descrever dados de autoridade.

As iniciativas descritas na Seção 4.2 foram selecionadas considerando sua abrangência, sua potencial contribuição para a proposição do modelo e a disponibilidade e acessibilidade de documentos descrevendo as etapas de seu desenvolvimento e/ou suas características.

1.5 Estrutura da pesquisa

No Capítulo 1, “Introdução”, foram apresentados o problema de pesquisa, a tese, a hipótese, os objetivos, a justificativa, a relevância, a motivação e a metodologia desta pesquisa. Os resultados parciais desta pesquisa são apresentados nos capítulos seguintes.

O Capítulo 2, “Catálogos e controle de autoridade”, aborda o controle de autoridade e os conceitos a ele relacionados (pontos de acesso, dados, registros, arquivos e sistemas de autoridade), bem como descreve o desenvolvimento dos dados de autoridade nos catálogos em livros, em fichas e digitais. Por fim, esse capítulo apresenta os modelos conceituais FRAD e FRSAD como síntese da compreensão atual acerca dos dados de autoridade no domínio bibliográfico.

O Capítulo 3, “Web Semântica, suas tecnologias e *Linked Data*”, introduz o conceito de Web Semântica, apresenta os princípios relacionados à publicação de *Linked Data* e descreve algumas das tecnologias relacionadas a esses princípios (URI, RDF, RDFS e OWL).

O Capítulo 4, “*Linked Data* no domínio bibliográfico: os dados de autoridade em foco”, descreve vocabulários que podem ser utilizados na publicação de dados de autoridade (SKOS, MADS/RDF e *RDA Element Sets*) e iniciativas de instituições que buscam publicar seus dados de autoridade como *Linked Data* (*Library of Congress Linked Data Service*, *datos.bne.es*, *data.bnf.fr*, VIAF e AGROVOC). Por fim, esse capítulo sintetiza alguns dos benefícios da publicação de dados de autoridade no domínio bibliográfico e além dele.

O Capítulo 5, “Modelo para a publicação de dados de autoridade como *Linked Data*”, inclui a apresentação das etapas do modelo proposto nesta pesquisa: planejamento; modelagem e mapeamento dos dados; tratamento, relacionamento e conversão; publicação; e *feedback* e retroalimentação.

Por fim, no Capítulo 6 são apresentadas as considerações finais sobre os resultados alcançados e sobre o modelo proposto nesta pesquisa.

2 Catálogos e controle de autoridade¹

Para os propósitos desta pesquisa, o termo catalogação adquire três significados. O primeiro deles, **catalogação enquanto processo**, refere-se à atividade de construção e de gestão de catálogos, principalmente no âmbito das bibliotecas. O segundo significado, **catalogação enquanto coisa**, refere-se às representações ou registros que integram a catalogação enquanto processo ou resultam dela. E o terceiro, **Catalogação enquanto (sub)disciplina** – grafado aqui com a inicial maiúscula – apresenta-se como parte da Biblioteconomia e da Ciência da Informação referindo-se ao conjunto de conhecimentos acerca das teorias, dos instrumentos de representação e das tecnologias relacionadas à catalogação enquanto processo e coisa.²

No entendimento da Catalogação, deve-se considerar que a catalogação enquanto processo antecede à Catalogação enquanto disciplina. Índícios da catalogação enquanto processo são notados já na antiguidade, quando bibliotecas como as de Nínive e de Alexandria representavam os recursos informacionais³ disponíveis em seus acervos (GARRIDO ARILLA, 1996, p. 62; SANTOS; PEREIRA, 2014, p. 16).

Durante a Idade Média, a catalogação enquanto processo é notada através dos catálogos dos mosteiros de Saint Riquier, na França, de Bobbio, na Itália, e de Lorsch, na Alemanha, do convento de St. Martin, na Inglaterra, e do *Trinity Hall*, em Cambridge, também na Inglaterra (SANTOS; PEREIRA, 2014, p. 16-18). Com o Renascimento tem-se o primeiro catálogo bibliográfico impresso, de responsabilidade de Aldo Manuzio, e os catálogos de Johann Trithem, Conrad Gesner e Andrew Maunsell (SANTOS; PEREIRA, 2014, p. 16-18; TAYLOR; JOUDREY, 2009, p. 71).

¹ O conteúdo deste capítulo está parcialmente apresentado nos artigos “O controle de autoridade no domínio bibliográfico: os catálogos em livros e em fichas” (ASSUMPÇÃO; SANTOS; ZAFALON, 2017b) e “O controle de autoridade no domínio bibliográfico: os catálogos digitais” (ASSUMPÇÃO; SANTOS; ZAFALON, 2017a).

² Uma distinção de natureza semelhante é realizada por Ortega (2011) e por Sousa, Saldanha e Tolentino (2017): “Adotamos a forma ‘Catalogação’ para indicar a área que se ocupa dos princípios e métodos de produção de bases de dados bibliográficas (termo mais amplo que catálogos de bibliotecas) e ‘catalogação’ para tratar das operações de produção dessas bases de dados.” (ORTEGA, 2011, p. 45) e “Catalogação iniciada com a letra ‘C’ maiúscula é empregada para designar a área de pesquisa e, com a letra ‘c’ minúscula refere-se ao processo, ou seja, ao ato de catalogar.” (SOUSA; SALDANHA; TOLENTINO, 2017, p. 332).

³ O termo “recurso informacional” é utilizado nesta pesquisa como termo genérico referente a um documento de um tipo qualquer. O termo “recurso”, utilizado a partir do Capítulo 3, é tratado como um termo genérico que inclui qualquer tipo de recurso, seja ele um recurso informacional ou não, como é caso de pessoas, instituições e conceitos, conforme descrito no Capítulo 3.

Sobre a presença dos catálogos e da catalogação enquanto processo na história das bibliotecas, Alves e Santos (2013, p. 23-24) apontam que,

Na construção e preservação do conhecimento humano, a mera acumulação dos acervos bibliográficos deixou de ser suficiente. O tratamento e a representação do acervo passaram a ser requeridos, demandando a criação de técnicas, de códigos, de padrões e de competências muito especializadas, enfim, um conjunto de processos e de procedimentos que não ocorreram e não ocorrem isoladamente.

A partir da segunda metade do século XIX, o desenvolvimento da Biblioteconomia e, já no século XX, da Ciência da Informação, contou com o surgimento e o desenvolvimento da Catalogação, acompanhados, obviamente, por avanços na catalogação enquanto processo. O desenvolvimento da Catalogação, devido às questões de ordem teórica e prática, foi marcado pela compreensão de que essa disciplina, junto de seus instrumentos, processos e produtos, poderia ser organizada a partir de duas principais abordagens: o tratamento descritivo e o tratamento temático, sendo “[...] que a distinção entre tais abordagens reside na busca do *o que* (materialização) [tratamento descritivo] e do *sobre o que* (teor) [tratamento temático] que convivem no âmbito do documento” (GUIMARÃES, 2009, p. 105).

Para os propósitos desta pesquisa, entende-se que o tratamento descritivo da informação, daqui em diante mencionado como Catalogação descritiva: (1) inclui os processos relacionados às descrições bibliográficas, aos pontos de acesso de responsabilidade e de títulos e, tradicionalmente, às informações administrativas sobre as representações (registros bibliográficos) dos recursos informacionais; (2) tem entre seus instrumentos as regras e as convenções de catalogação, apresentados principalmente na forma de códigos e de manuais, e os padrões de metadados; e (3) tem como principal produto os registros bibliográficos, ainda que nestes sejam incluídos também produtos do tratamento temático.

Para Barité (1998, p. 124⁴ apud GUIMARÃES, 2009, p. 106), o tratamento temático da informação, centra-se nas questões relativas “à análise, descrição e representação do conteúdo dos documentos, bem como suas inevitáveis interfaces com as teorias e sistemas de armazenamento e recuperação da informação”. O tratamento temático da informação é identificado na literatura sob três vertentes teóricas: “a catalogação de

⁴ BARITÉ, M. **Referenciales teóricos vigentes en el área de tratamiento temático de la información y su expresión metodológica**. Porto Alegre: ABEED, 1998.

assunto (*subject cataloguing*) de matriz norte-americana, a indexação (*indexing*) de matriz inglesa e a análise documental (*analyse documentaire*), de matriz francesa” (GUIMARÃES, 2009, p. 106).

O catálogo, às vezes chamado de catálogo bibliográfico ou arquivo bibliográfico, é o principal resultado tangível dos processos descritivos e temáticos da catalogação realizados nas bibliotecas. No domínio bibliográfico, os catálogos são entendidos como ferramentas para o acesso aos recursos informacionais de uma instituição, como meios de comunicação entre os recursos informacionais e seus usuários (MEY; SILVEIRA, 2009, p. 12) e como conjuntos organizados de registros bibliográficos (TAYLOR, 2006, p. 6). As unidades básicas que compõem os catálogos, os chamados registros bibliográficos,

Apresentam elementos simbólicos aqui denominados de informações documentárias - os quais possuem competência para informar, pois operam com a significação. Informação documentária é aquela apreendida, registrada e armazenada em um sistema de informação (documentária) de forma a ser passível de recuperação e uso para os mais diversos fins demandados pela sociedade. (ORTEGA; LARA, 2010, p. 9).

Para a Declaração dos Princípios Internacionais de Catalogação, um registro bibliográfico é “um conjunto de elementos de dados que descreve e provê acesso a um recurso bibliográfico e identifica obras e expressões relacionadas” (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009, p. 10, tradução nossa). Tradicionalmente, um registro bibliográfico é composto por uma descrição bibliográfica, “um conjunto de dados bibliográficos que identifica um recurso bibliográfico” (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009, p. 9, tradução nossa), e por um ou mais pontos de acesso, “um nome, termo, código, etc. por meio do qual dados bibliográficos ou de autoridade são buscados e identificados” (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009, p. 9, tradução nossa). Descrições bibliográficas e pontos de acesso resultam dos processos de catalogação. Dados como o título, as indicações de responsabilidade, as pessoas relacionadas, a data de publicação e o número de páginas proveem de processos descritivos, enquanto dados como os pontos de acesso de assunto e notações de classificação temática proveem de processos temáticos.

Ainda que o catálogo com seus registros bibliográficos seja o principal objeto de estudo da Catalogação, deve ser observado que outras ferramentas de acesso à informação com outros tipos de registros e de dados também são de interesse desta

disciplina. Entre essas ferramentas, têm-se os arquivos de autoridade, que, por meio de dados de autoridade, prestam-se principalmente ao controle dos pontos de acesso utilizados nos catálogos.

Os dados de autoridades são os principais objetos de estudo desta pesquisa. Para uma explanação acerca desses dados, de sua função nos catálogos e de seu desenvolvimento ao longo da história da Catalogação, nas seções deste capítulo são abordados o controle de autoridade e os conceitos a ele relacionados (Seção 2.1), a função dos dados de autoridade nos catálogos analógicos (Seção 2.2) e digitais (Seção 2.3) e os modelos conceituais sobre dados de autoridade publicados na última década (Seção 2.4), provendo, assim, parte do embasamento teórico necessário ao modelo para publicação de dados de autoridade como Linked Data proposto no Capítulo 5.

2.1 Controle de autoridade

Como destacado, os tratamentos descritivo e temático voltam-se à materialização e ao teor dos recursos informacionais (GUIMARÃES, 2009, p. 105) e, a partir destes, são construídas representações desse teor e dessa materialização. Nos catálogos de bibliotecas, essas representações constituem os registros bibliográficos, que passam a incluir, portanto, dados resultantes do tratamento descritivo e do tratamento temático, entre outros tipos de dados⁵.

Resultantes do tratamento descritivo dos recursos informacionais, tem-se dados como: título principal, títulos variantes e outras informações sobre o título; indicações relacionando pessoas e entidades coletivas associadas ao recurso informacional; local, data e responsáveis pela publicação, produção, distribuição ou fabricação; formato, dimensão e extensão do suporte, etc. Enquanto que do tratamento temático resultam dados como os termos e as notações que representam os conceitos abordados no teor do recurso informacional.

De modo geral, nos catálogos digitais atuais qualquer dado presente em um registro bibliográfico pode ser utilizado para sua recuperação, por exemplo, a data de publicação ou o idioma do recurso informacional. Ao longo da história da Catalogação,

⁵ Entre esses outros tipos de dados estão os dados administrativos ou técnicos acerca dos recursos informacionais e dos registros bibliográficos. Estes tipos de dados não são tratados nesta seção por não apresentarem contribuição relevante para a compreensão do conceito de controle de autoridade.

no entanto, restrições de ordem prática e tecnológica limitaram quais dados seriam utilizados para a recuperação dos registros bibliográficos nos catálogos. Assim, de modo geral, os dados escolhidos foram o título do recurso informacional, os nomes dos responsáveis por ele e os termos referentes aos conceitos tratados em seu conteúdo, como notado na segunda metade do século XIX e no início do século XX, descritos na Seção 2.2.

A utilização desses dados, principalmente dos nomes e dos termos, como eram apresentados nos recursos informacionais, por exemplo, o nome do autor como indicado na página de rosto de um livro ou o termo utilizado pelo autor para referir-se a um conceito⁶ no decorrer do texto, poderia causar problemas na recuperação dos registros bibliográficos, uma vez que os responsáveis e os conceitos, aqui generalizados sob o termo entidades,⁷ não mantêm relações de univocidade com seus nomes, visto que:

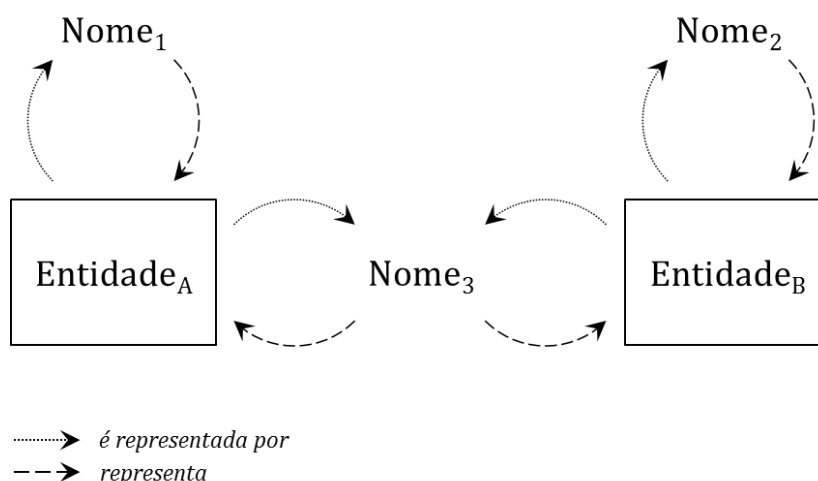
- **uma entidade poderia ser representada por mais de um nome**, tanto nos recursos informacionais, quanto no conjunto de conhecimentos do usuário do catálogo, por exemplo:
 - uma pessoa poderia ser representada por seu nome completo, nome abreviado, nome anterior ou posterior ao casamento, etc.;
 - uma entidade coletiva poderia ser representada por seu nome oficial, nome pelo qual é mais conhecida, nome anterior, nome mais recente, sigla, nome em outro idioma ou alfabeto, etc.;
 - um conceito poderia ser representado por um nome científico, popular ou local, uma sigla, por variações lexicais, etc.
- **um nome poderia representar mais de uma entidade**, tanto nos recursos informacionais, quanto no conjunto de conhecimentos do usuário do catálogo; por exemplo, um mesmo nome pode ser utilizado por diversas pessoas ou entidades coletivas e um mesmo termo pode referir-se a diferentes conceitos em diferentes áreas do conhecimento.

⁶ “Um conceito é definido como uma ‘unidade de pensamento’, enquanto o termo é um ‘palavra ou frase usada para rotular um conceito’ (ISO, 2013).” (VAN HOOLAND; VERBORGH, 2015, p. 3, tradução nossa).

⁷ Uma entidade é entendida aqui como “Algo com caráter unitário e autocontido; algo com existência independente ou separada; uma abstração, um conceito ideal, um objeto de pensamento um ou objeto transcendental” (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009, p. 11, tradução nossa).

Essas duas situações que representam as relações entre as entidades e seus nomes estão esquematizadas na Figura 1, onde *Entidade_A* é representada por *Nome₁* e *Nome₃*; *Entidade_B* é representada por *Nome₂* e *Nome₃*; e *Nome₃* representa ambas as *Entidade_A* e *Entidade_B*.

Figura 1 – Relações entre as entidades e seus nomes



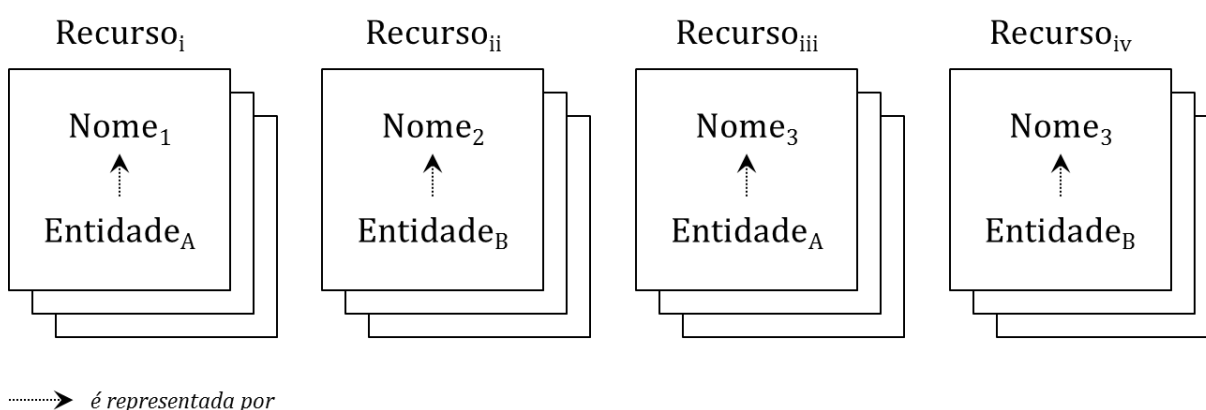
Fonte: Elaborada pelo autor.

Essas duas situações dificultariam a recuperação dos registros bibliográficos nos catálogos, pois (1) resultariam na recuperação de registros bibliográficos irrelevantes ao usuário, nos casos em que um nome representasse mais de uma entidade, e (2) impediriam a recuperação de todos os registros bibliográficos associados a uma entidade, por exemplo, todos os registros dos livros de um mesmo autor ou sobre um mesmo conceito, nos casos em que uma entidade fosse representada por mais de um nome. Os problemas de recuperação decorrentes dessas situações estão esquematizadas na Figura 2, sendo que:

- um usuário que deseja obter recursos informacionais associados à *Entidade_A*, ao realizar uma busca utilizando o *Nome₃* recupera os recursos informacionais em que a *Entidade_A* é representada pelo *Nome₃* (*Recurso_{iii}*) e também os recursos informacionais em que o *Nome₃* representa outras entidades, no caso, a *Entidade_B* (*Recurso_{iv}*), portanto, recursos informacionais possivelmente irrelevantes ao usuário;

- um usuário que deseja obter todos os recursos informacionais associados à *Entidade_B*, ao realizar uma busca utilizando o *Nome₂* recupera somente os recursos informacionais em que a *Entidade_B* é representada pelo *Nome₂* (*Recurso_{ii}*), excluindo, assim, os recursos informacionais em que a *Entidade_B* é representada por outros nomes (*Recurso_{iv}*).

Figura 2 – Relações entre entidades, nomes e recursos



Fonte: Elaborada pelo autor.

A busca por soluções para os problemas de recuperação nos catálogos decorrentes da falta de univocidade entre as entidades e seus nomes levou ao surgimento do conceito de **controle de autoridade**.

As definições de controle de autoridade encontradas na literatura, mesmo voltando-se para a necessidade de garantir a univocidade entre as entidades e os nomes utilizados para representá-las nos catálogos, apresentam diferentes entendimentos acerca do que se constitui como o controle de autoridade: alguns autores entendem o “controle de autoridade como um estado que se alcança por meio da realização de um conjunto de atividades”, enquanto outros entendem o controle de autoridade como “o próprio conjunto de atividades ou algumas atividades desse conjunto” (ASSUMPÇÃO; SANTOS, 2012, p. 5-6). Nos parágrafos seguintes são brevemente apresentados alguns autores que demonstram esses entendimentos.

Para Taylor (1984, p. 2 apud CLACK, 1988, p. 36, tradução nossa) o controle de autoridade é “o processo de manter a consistência dos cabeçalhos⁸ em um arquivo bibliográfico por meio de referências a um arquivo de autoridade”. Clack (1988, p. 35, tradução nossa) destaca que o controle de autoridade é um “[...] processo técnico executado em um arquivo bibliográfico para prover estrutura a esse arquivo”, sendo que esse processo envolve

[...] vários processos interdependentes: (1) a certificação de que o cabeçalho da entrada é único, (2) a aplicação de um conjunto rígido de padrões para a escolha da forma dos cabeçalhos das entradas, e (3) a criação de vínculos entre as entradas autorizadas e todas as variações dessas entradas. (CLACK, 1988, p. 35, tradução nossa).

A autora destaca também que a função do controle de autoridade é assegurar a unicidade e a consistência das formas dos nomes, dos títulos e dos assuntos utilizados como pontos de acesso nos registros bibliográficos, de modo a facilitar o acesso eficiente à informação nos catálogos (CLACK, 1988, p. 36).

Para Herrero Pascual (1999, p. 121, tradução nossa),

O controle de autoridade é o processo de unificar, mediante a utilização de uma forma padronizada, os pontos de acesso dos catálogos automatizados e também de apresentar as relações entre distintos pontos de acesso. Ou seja, significa a padronização dos nomes de pessoas, de entidades, títulos uniformes ou assuntos, que podem constituir um ponto de acesso principal ou os secundários de um catálogo automatizado.

Enquanto esses autores definem o controle de autoridade como um processo ou um conjunto de processos que visam à consistência dos pontos de acesso em um catálogo, outros consideram o controle de autoridade algo alcançado quando a consistência dos pontos de acesso é mantida por meio de um conjunto de processos, em outras palavras, consideram o controle de autoridade um fim e não um meio.

Além de entender que o controle de autoridade é alcançado com a consistência dos pontos de acesso utilizados em um catálogo, Burger (1985, p. 3, tradução nossa) aponta como necessário ao controle de autoridade o estabelecimento de políticas de catalogação que garantam tal consistência: “O controle de autoridade é alcançado quando os cabeçalhos utilizados em um catálogo estão consistentes e quando um

⁸ “Cabeçalho” é sinônimo de ponto de acesso. Essa denominação advém dos catálogos em fichas, em que os pontos de acesso eram incluídos no início da ficha catalográfica, como um cabeçalho de uma página.

mecanismo (o arquivo de autoridade) e políticas de catalogação relacionadas foram estabelecidos para assegurar essa consistência”.

Chan (1994, p. 3) também considera o controle de autoridade algo alcançado em um catálogo quando termos uniformes são utilizados para representar os nomes e os assuntos que servem como pontos de acesso para os registros bibliográficos, de modo que os registros bibliográficos associados a uma mesma entidade não se dispersem entre os sinônimos e as formas variantes dos nomes. Para tanto, a autora entende como atividades de controle de autoridade (1) a integração de registros de autoridade padronizados no sistema local e (2) a preparação de registros de autoridade para os nomes e os assuntos não disponíveis nos arquivos de autoridade utilizados como padrão, por exemplo, nos arquivos de autoridade de uma agência bibliográfica nacional ou programa de catalogação cooperativa (CHAN, 1994, p. 13-14).

Taylor e Joudrey (2009, p. 249, tradução nossa) definem o controle de autoridade como “[...] o resultado do processo de manutenção da consistência da forma verbal utilizada para representar um ponto de acesso e dos processos adicionais de apresentação dos relacionamentos entre nomes, obras e assuntos”.

Gorman (2004, p. 12, tradução nossa) enfatiza a necessidade do controle de autoridade ao considerá-lo, junto do controle bibliográfico, “dois lados de uma mesma moeda”. Para o autor,

No mínimo, o controle bibliográfico é literalmente impossível sem o controle de autoridade. A catalogação não pode existir sem pontos de acesso padronizados e o controle de autoridade é o mecanismo pelo qual alcançamos o grau necessário de padronização. A catalogação trata de ordem, lógica, objetividade, denotação precisa e consistência, e deve ter mecanismos para assegurar esses atributos. (GORMAN, 2004, p. 12, tradução nossa)

O controle de autoridade é central e vital às atividades que nós chamamos de catalogação. Catalogação – a reunião lógica de dados bibliográficos dentro de registros recuperáveis e utilizáveis – é a atividade que permite a biblioteca prosseguir em suas missões centrais de serviço e de acesso gratuito e aberto a todo o conhecimento e informação registrados. Não podemos ter um serviço real de biblioteca sem uma arquitetura bibliográfica e não podemos ter essa arquitetura bibliográfica sem o controle de autoridade. (GORMAN, 2004, p. 21, tradução nossa).

Sobre a importância do controle de autoridade, Assumpção e Santos (2012), ao analisarem as declarações de Cutter, de Verona, de Domanovszky, de Lubetzky e de Svenonius acerca dos objetivos e funções dos catálogos, assim como aquelas

apresentadas nos “Princípios de Paris”, no *Functional Requirements for Bibliographic Records* (FRBR) e na Declaração dos Princípios Internacionais de Catalogação, destacam que o controle de autoridade é necessário para que o catálogo permita ao usuário encontrar um recurso informacional conhecido, encontrar um conjunto de recursos informacionais que compartilham determinada característica e navegar no catálogo por meio dos relacionamentos existentes entre as entidades.

Alguns autores, ao definirem o controle de autoridade o relacionam ao conceito de **trabalho de autoridade**:

O termo *authority control* é utilizado frequentemente com um termo guarda-chuva para indicar um conjunto de práticas e procedimentos (*authority work*) que visa à criação de um “arquivo” (*authority file*) de registros de autoridade (*authority record*). (BOZZARELLI, 2004, p. 5, tradução nossa).

Segundo Marais (2004, p. 62, tradução nossa), o “controle de autoridade é um termo abrangente e genérico para os processos técnicos do trabalho de autoridade e objetiva alcançar o controle sobre as formas variantes dos pontos de acesso”. Para Burger (1985, p. 3, tradução nossa), o “trabalho de autoridade permite que o controle de autoridade ocorra”. O autor, na falta de uma definição abrangente sobre o escopo e o propósito, opta por inferir, a partir da literatura, que trabalho de autoridade consiste, ao menos, em cinco processos:

1. a criação de *registros de autoridade* (que, por sua vez, são utilizados para criar registros bibliográficos autorizados);
2. a compilação de registros em um *arquivo de autoridade*;
3. a vinculação do arquivo de autoridade ao arquivo bibliográfico; juntos eles formam um *sistema de autoridade*;
4. a manutenção do arquivo e do sistema de autoridade;
5. a avaliação do arquivo e do sistema de autoridade. (BURGER, 1985, p. 3, grifo do autor, tradução nossa).

Assumpção e Santos (2012, p. 6), concordam com Burger e enfatizam que a consistência necessária ao controle de autoridade implica “[...] uma entidade ser identificada por apenas um ponto de acesso autorizado e um ponto de acesso autorizado identificar somente uma entidade”, o que é destacado por Chan (2007, p. 164-165) com o princípio dos cabeçalhos uniformes e o princípio dos cabeçalhos únicos.

Hagler (1997, p. 112-113, tradução nossa) considera que as tarefas de

1) determinar se existe ou não um relacionamento significativo entre diferentes nomes ou termos e se um mesmo nome ou termo pode referir-se a mais de uma diferente pessoa, conceito, etc.

2) estabelecer e vincular todos os possíveis pontos de acesso que podem razoavelmente expressar cada pessoa, conceito, etc. distinto

são coletivamente chamadas de trabalho de autoridade. Ainda segundo o autor, o trabalho de autoridade impõe a decisão do catalogador sobre como melhor organizar e apresentar as entidades e os significados implícitos nas palavras em linguagem natural e que, portanto, tal trabalho não pode ser reduzido à mera aplicação de regras (HAGLER, 1997, p. 113).

Maxwell (2002, p. 3-4, destaque do autor, tradução nossa) ressalta a importância do trabalho de autoridade enquanto responsável por manter registros das decisões tomadas em relação a cada ponto acesso autorizado utilizado no catálogo:

Geralmente há espaço para a decisão do catalogador na escolha da forma para um determinado nome ou assunto, assim, diferentes catalogadores podem chegar a diferentes cabeçalhos para um mesmo nome. [...] Se um registro das decisões anteriores não é mantido, o único modo de descobrir [qual decisão foi tomada] será verificando os registros bibliográficos e observando o que foi utilizado no passado. Isso pode funcionar na maioria dos casos, mas um método mais eficiente é manter separado um registro das decisões tomadas, de modo que os catalogadores não precisem retomá-las cada vez que um cabeçalho for necessário. Esses registros criam uma base de dados ou um arquivo usualmente referido como um “arquivo de autoridade”.

Os registros aos quais Maxwell refere-se são hoje denominados **registros de autoridade**. Segundo a Declaração dos Princípios Internacionais de Catalogação, um registro de autoridade é “um conjunto de elementos de dados que identifica uma entidade e pode ser utilizado para facilitar o acesso ao ponto de acesso autorizado para a entidade ou a exibição de qualquer ponto de acesso para ela” (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2009, p. 9, tradução nossa). Taylor e Joudrey (2009, p. 252, tradução nossa) acrescentam que um registro de autoridade “inclui evidências de todas as decisões tomadas e de todos os relacionamentos entre as variações identificadas”.

Os dados dos registros de autoridade são considerados **dados de autoridade**. Segundo o modelo conceitual *Functional Requirements for Authority Data* (FRAD) (IFLA WORKING GROUP ON FRANAR, 2013, p. 1), os dados de autoridade representam os pontos de acesso controlados e outras informações que as instituições utilizam para agrupar as obras de uma pessoa, família ou entidade coletiva, ou as várias edições de uma obra.

Para os propósitos desta pesquisa, considera-se que o termo dados de autoridade indica o conjunto de dados formado pelos pontos de acesso controlados utilizados na identificação de uma determinada entidade e pelos dados que apoiam o uso destes pontos de acesso, sejam eles dados sobre a entidade, por exemplo, pontos de acesso relacionados, dados biográficos ou históricos, ou sobre o processo de estabelecimento dos pontos de acesso, por exemplo, as regras utilizadas, as fontes de informação consultadas e a instituição na qual ou para a qual os dados de autoridade foram criados.

Esta compreensão acerca dos dados de autoridade respalda-se em conceitos de dados como o apresentado por Santos e Sant'Ana (2013, p. 205):

[...] dado é uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

Os tipos de dados de autoridade variam conforme (1) os instrumentos de descrição, por exemplo, regras de catalogação e padrões de metadados, utilizados na criação do registro de autoridade, e (2) a natureza da entidade representada pelo registro, por exemplo, os dados de autoridade associados a uma pessoa são diferentes dos dados associados a uma entidade coletiva que, por sua vez, são diferentes dos dados associados a um conceito, etc. Uma relação dos principais tipos de dados de autoridade é apresentada na Seção 2.4, em que são abordados os modelos conceituais FRAD e *Functional Requirements for Subject Authority Data* (FRSAD).

Os registros de autoridade, quando reunidos, constituem os **arquivos de autoridade** (BURGER, 1985, p. 3; TAYLOR, 2006, p. 19-20), também denominados catálogos de identidade (MEY; SILVEIRA, 2009, p. 195). Para Hagler (1997, p. 117), o arquivo de autoridade é o principal produto do trabalho de autoridade e a expressão tangível de um vocabulário controlado. Segundo o autor, esse arquivo nunca está finalizado, pois “novos nomes, termos e relacionamentos devem ser acrescentados e os obsoletos devem ser removidos, e o trabalho já realizado está sujeito a melhorias” (HAGLER, 1997, p. 118, tradução nossa).

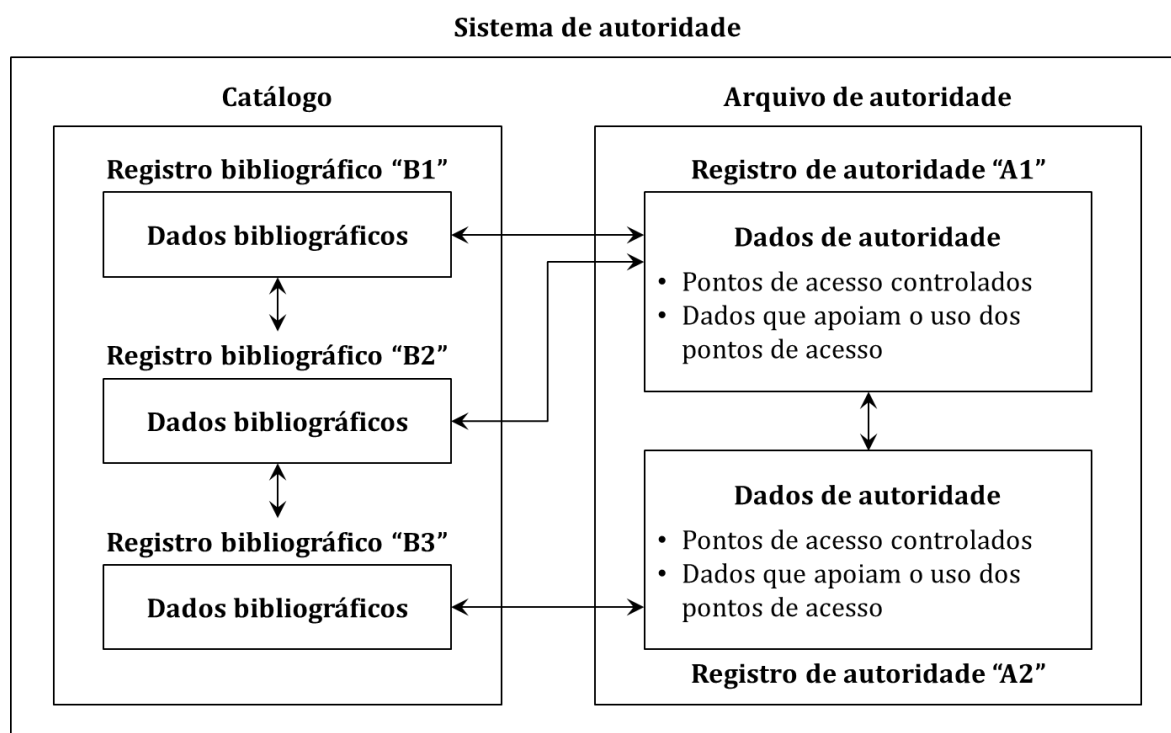
Calhoun (1998, tradução nossa) lista quatro funções dos arquivos de autoridade: (1) “função de autoridade”, relativa à manutenção da consistência nos catálogos; (2) “função de localização”, relativa à habilidade dos usuários buscarem e encontrarem entidades utilizando um de seus vários nomes; (3) “função informativa”, relativa à

documentação das fontes de informação consultadas pelo catalogador para a criação dos pontos de acesso; e (4) “função de manutenção”, relativa à possibilidade de uso do arquivo de autoridade para a identificação e a correção de erros no catálogo.

Registros e arquivos de autoridade podem apresentar-se de diferentes formas em razão das tecnologias utilizadas e das políticas de catalogação de cada instituição. Em sistemas manuais, por exemplo, registros representando nomes de pessoas e de entidades coletivas e registros representando conceitos eram frequentemente mantidos em arquivos de autoridade separados, enquanto que em sistemas digitais podem ser mantidos em uma mesma base de dados e integralmente vinculados ou não ao catálogo (CHAN, 1994, p. 12).

A junção de um arquivo de autoridade com um catálogo, segundo Burger (1985), resulta em um **sistema de autoridade**, cujos componentes e relacionamentos estão sintetizados na Figura 3. Para o autor, o fato de arquivo de autoridade conter os pontos de acesso utilizados no catálogo já indica um vínculo implícito entre os dois, no entanto, a explicitação deste vínculo é que indica o tipo de relacionamento entre o catálogo e o arquivo de autoridade (BURGER, 1985, p. 32).

Figura 3 – Sistema de autoridade com seus componentes e relacionamentos



Na década de 1980, Burger identificou quatro tipos possíveis para esse relacionamento (vínculo): (1) o arquivo de autoridade independente do catálogo, quando os dois existem como bases de dados distintas, sem que as alterações realizadas no arquivo de autoridade sejam refletidas no catálogo; (2) o arquivo de autoridade coordenado com o catálogo, quando não há um vínculo explícito entre os dois, mas as mudanças realizadas no arquivo de autoridade são realizadas também no catálogo, como ocorre com os catálogos em fichas; (3) o arquivo de autoridade como parte do catálogo, quando não há registros de autoridade e os registros bibliográficos do catálogo são utilizados para o registro e a consulta aos pontos de acesso autorizados; e (4) o arquivo de autoridade conectado ao catálogo, quando os dois estão vinculados de modo que uma alteração realizada no arquivo de autoridade seja automaticamente refletida no catálogo (BURGER, 1985, p. 32-34). Alguns desses vínculos são notados na revisão sobre o controle de autoridade nos catálogos em livros, em fichas e digitais, conduzida nas Seções 2.2 e 2.3.

O trabalho envolvido no controle de autoridade, por demandar pessoal especializado e tempo, é custoso (MAXWELL, 2002, p. 6; MARAIS, 2004, p. 70), o que faz com que, frequentemente, a busca por palavras-chave não controladas (*keyword searching*) seja considerada suficiente para a recuperação de recursos informacionais (TAYLOR; JOUDREY, 2009, p. 189). Nesse sentido, em defesa do controle de autoridade, Taylor e Joudrey (2009, p. 189, tradução nossa) destacam que

[...] pesquisas têm demonstrado que a busca por palavras-chave pode gerar resultados falsos (ou seja, recuperar resultados irrelevantes) porque a palavra recuperada possui um significado diferente do pretendido e pode também reduzir a revocação, pois os sinônimos e os quase sinônimos não foram recuperados com a palavra pesquisada.

Tillett (2002, p. 7, tradução nossa) defende o controle de autoridade destacando sua importância enquanto diferencial de um catálogo em relação a uma mera lista:

Nos catálogos bibliográficos automatizados e, atualmente, na Web – assim como nos antigos catálogos manuais em forma de livro ou de fichas – o conceito de autoridade segue sendo essencial para assegurar a estrutura sindética que transforma uma simples lista de busca em um autêntico catálogo.

Também nesse sentido, Taylor (2006, p. 20) pontua que o controle de autoridade, por permitir o relacionamento entre as obras, faz com que a catalogação seja mais que

um processo de criar registros bibliográficos para representar obras sem qualquer relacionamento aparente entre si.

Por fim, como síntese desta seção, para os propósitos desta pesquisa, entende-se que o controle de autoridade é um estado alcançado em um catálogo quando os pontos de acesso são utilizados consistentemente, ou seja, quando um ponto de acesso autorizado representa uma única entidade e quando uma entidade é representada por apenas um ponto de acesso autorizado. O controle de autoridade é alcançado mediante os procedimentos agrupados sob o termo trabalho de autoridade; entre estes procedimentos estão a criação de dados de autoridade a partir da definição de pontos de acesso autorizados, não autorizados e relacionados com vistas à criação e a gestão de registros, arquivos e sistemas de autoridade (ASSUMPCÃO; SANTOS, 2012, p. 6).

A necessidade do controle dos pontos de acesso nos catálogos e a existência de instrumentos para esse controle são observadas no decorrer da história da Catalogação, ora de modo implícito, ora de modo explícito. Como pontuam Alves e Santos (2013, p. 23-24), “[...] as bibliotecas procuraram melhorar, cada vez mais, as formas de construção dos catálogos com o desenvolvimento e a utilização de métodos eficazes e econômicos, sempre utilizando as tecnologias disponíveis”. A busca por essa melhoria e o uso constante de tecnologias fizeram com que os catálogos evoluíssem dos tablets de argila até os bancos de dados bibliográficos atuais, passando, entre outros, pelos catálogos em livros, pelos catálogos em fichas e pelas fitas magnéticas e formatos de intercâmbio de dados desenvolvidos ainda nas décadas de 1960 e 1970. Evoluindo junto dos catálogos esteve a necessidade de controle dos pontos de acesso e os instrumentos utilizados para tanto, como é apresentado nas duas seções seguintes deste capítulo, que abordam o controle de autoridade nos catálogos em livros e em fichas (Seção 2.2) e nos catálogos digitais (Seção 2.3).

2.2 Controle de autoridade nos catálogos em livros e em fichas

Ainda que a existência de catálogos remonte a períodos da antiguidade (SANTOS; PEREIRA, 2014, p. 15), esta revisão de literatura inicia-se com considerações acerca dos catálogos e do controle de autoridade a partir, aproximadamente, da segunda metade do século XIX. Essa delimitação respalda-se em dois principais motivos. Primeiramente, nota-se a escassez de informações sobre os catálogos anteriores a esse período, escassez

essa decorrente da falta de acesso às fontes de informação primárias (por exemplo, os catálogos e as regras de catalogação anteriores ao século XIX), e das revisões de literatura e pesquisas históricas que não proveem informações suficientes para uma análise sobre os dados presentes nesses catálogos.

Em segundo lugar, tem-se que a partir da segunda metade do século XIX é que os princípios de catalogação receberam uma primeira sistematização (ORTEGA; LARA, 2010, p. 10) e tiveram início “as tentativas de consolidação de princípios e de regras unificadas para a representação, no intuito de criar um embasamento teórico para as práticas de descrição” (ALVES; SANTOS, 2013, p. 27). Também data dessa época o início do período semimecanizado da catalogação (ALVES; SANTOS, 2013, p. 27-28) tendo entre seus expoentes “Antony Panizzi, Charles Coffin Jewett e Charles Ammi Cutter com a revisão e o aprimoramento das regras de catalogação existentes na época, com o estabelecimento de novas regras e, principalmente, com o estabelecimento de princípios para o processo de catalogação” (ALVES; SANTOS, 2013, p. 27-28).

Segundo Obille (2012, p. 7), a partir da segunda metade do século XIX surgiram objetivos que se tornaram subjacentes aos catálogos e aos instrumentos utilizados em sua construção. Esses objetivos, criados a partir de questões de natureza teórica e não apenas prática, são considerados parte do paradigma vigente ainda hoje na Catalogação (OBILLE, 2012, p. 7).

É nesse período que, segundo Tait (1969⁹ apud GUERRINI, 2009, p. 1, tradução nossa), “se inicia a introdução do conceito de responsabilidade intelectual por parte de Antony Panizzi, aperfeiçoado por Charles Ammi Cutter e continuado ininterruptamente até os dias atuais”.

Desse período remontam os esforços de Panizzi, que, no *British Museum*, advogou a favor de um compêndio de regras para a elaboração do catálogo da biblioteca do referido museu (LUBETZKY, 2000; SANTOS; PEREIRA, 2014, p. 24-25). O compêndio de regras, publicado em 1841 no início do primeiro volume do *Catalogue of Printed Books in the British Museum*, tornou-se conhecido como as 91 regras de Panizzi e é considerado um dos primeiros códigos de catalogação de que se tem conhecimento (BARBOSA, 1978, p. 26; GARRIDO ARILLA, 1996, p. 76).

⁹ TAIT, J. A. **Authors and titles**: an analytical study of the author concept in codes of cataloguing rules in the English language, from that of the British Museum in 1841 to the Anglo-American Cataloguing Rules 1967. London: Bingley, 1969.

Segundo Lubetzky (2000, p. 425, tradução nossa), o estabelecimento das 91 regras de Panizzi foi parte do primeiro acontecimento pivô da história da catalogação anglo-americana, uma vez que ele

Diz respeito principalmente ao fundamento ideológico do catálogo – o problema fundamental de quais seriam os objetivos aos quais o catálogo de uma biblioteca deveria ser projetado para servir e os métodos para alcançar esses objetivos. E influenciou as discussões e o curso da catalogação anglo-americana desde então [...].

Ainda segundo Lubetzky, as 91 regras de Panizzi diferem das regras até então utilizadas principalmente no entendimento do que um livro¹⁰ representa:

Como as regras de Panizzi diferem das regras previamente utilizadas no *British Museum* e em outros lugares?

Curiosamente, elas diferem principalmente, e basicamente, na visão do que um livro representa. As regras tradicionais veem os livros de uma biblioteca como entidades distintas e o catálogo como um registro designado a ajudar alguém a encontrar o livro desejado sob o nome do autor, o título ou sob qualquer outra indicação pela qual o usuário pode razoavelmente buscá-lo. As regras de Panizzi compreendiam os *livros* como edições de *obras* distintas, sendo os livros que representassem a mesma obra integrados e arranjados no catálogo em uma ordem prescrita, de modo que o leitor que recorresse ao catálogo da biblioteca para buscar um determinado livro o encontraria em um contexto com outros livros representando edições da mesma obra e poderia selecionar a edição ou as edições que melhor servissem aos seus interesses. (LUBETZKY, 2000, p. 422, tradução nossa).

Na época de Panizzi, os registros dos catálogos eram escritos à mão, datilografados ou impressos (CHAN, 2007, p. 22) e apresentados principalmente em fichas e em páginas agrupadas no formato de códices, os chamados catálogos em livro (*book catalogs*) (SVENONIUS, 2000, p. 62; TILLET, 1992, p. 24). Nesses catálogos, o modo com que as entradas eram arranjadas determinava como elas seriam recuperadas e apresentadas (CHAN, 2007, p. 24).

A tecnologia utilizada na criação dos catálogos em livros – o códice – influenciou diretamente as 91 regras de Panizzi, como destaca Tillett (1989, p. 152, tradução nossa):

Como resultado das restrições econômicas sobre o tamanho dos catálogos em livro e o extenso tempo envolvido na redação de fichas para catálogos escritos a mão, as regras de Panizzi instruíram para que um item bibliográfico fosse descrito completamente uma única vez por meio de uma “entrada”.

¹⁰ É necessário considerar que o termo “livro” tem sido usado na catalogação, muitas vezes, como termo abrangente para designar os diversos tipos de documentos presentes nos acervos das bibliotecas.

Garrido Arilla (1996, p. 76) aponta que as 91 regras foram redigidas para a criação de catálogos alfabéticos, principalmente de autores pessoais ou corporativos, que permitissem aos usuários a localização rápida e fácil de uma obra ou o agrupamento das obras de um mesmo autor com suas distintas edições e traduções. A organização das entradas nos catálogos em livros de acordo com as 91 regras é descrita por Svenonius (2000, p. 62, destaque do autor, tradução nossa):

Abaixo do nome de cada autor eram listadas, por título, as obras escritas por ele. A primeira edição de uma obra custodiada pela biblioteca era descrita por completo. Se a biblioteca possuísse uma segunda edição, sua entrada seria incluída abaixo da entrada da primeira edição assim: “----- Another edition.” Se necessárias, informações que serviam para distinguir a segunda edição da primeira seriam providas. Se a biblioteca possuísse mais de uma cópia da edição, era descrito: “-----
- Another copy.”

Para a autora, essa forma de organização das entradas

[...] assegurava que cada documento seria integrado ao catálogo no sentido de que seria apresentado no contexto de todas as edições, traduções e adaptações da obra manifestada nele. [...] A exibição hierárquica não só assegurava a integração; ela também alcançava economia de expressão ao limitar o que tinha de ser dito sobre uma entrada a somente aquelas características que não poderiam ser derivadas usando relacionamentos de participação e de inclusão. (SVENONIUS, 2000, p. 62, tradução nossa).

Tillett (1989, p. 152) destaca que Panizzi incluiu remissivas para prover acesso mais completo às entradas e para tornar o catálogo mais que uma mera lista ou inventário da coleção. Sobre a função das remissivas, Svenonius (2000, p. 62-63, tradução nossa) explica que elas indicavam os relacionamentos não hierárquicos entre os recursos informacionais:

Nos catálogos em livros, os relacionamentos de um tipo não hierárquico (ou seja, outros que não fossem de participação ou de inclusão) eram indicados por remissivas. Remissivas Ver vinculavam obras relacionadas, tal como a Bíblia e um comentário sobre ela. Elas eram usadas também para alcançar a disposição de obras quando o acesso era através de atributos secundários. Por exemplo, se um documento era de autoria de duas pessoas, a descrição principal (ou entrada principal) seria feita para o primeiro autor, e sob o nome do autor secundário uma remissiva Ver direcionaria o usuário para a descrição principal, que exibiria o documento em seu contexto integrado.

Svenonius (2000, p. 62-63, tradução nossa) complementa que “a necessidade de remissivas bibliográficas de um tipo não hierárquico para auxiliar na navegação e na

disposição foi bem entendida por Panizzi, que observou que quanto maior a quantidade de remissivas, mais útil o catálogo seria”.

Em suas 91 regras, Panizzi (1841, p. vii) especificou o uso de três tipos de remissivas: de nome para nome, de nome para obra e de obra para obra. Tillett (1989, p. 152; 1992, p. 24) explica esses três tipos de remissivas:

- nome para nome: direcionava de uma forma variante do nome para a forma do nome escolhida para a entrada;
- nome para obra: direcionava para a entrada de uma obra a partir de cabeçalhos que poderiam ser considerados tão importantes quanto o cabeçalho presente na entrada da obra; esses cabeçalhos que direcionavam para a entrada da obra incluíam nomes pessoais, de entidades coletivas, de conferências, de localizações geográficas e de obras (títulos);
- obra para obra: direcionava de uma obra para outra, comumente de partes de uma obra para a obra como um todo.

Enquanto Panizzi publicava suas regras na Inglaterra, Charles Coffin Jewett iniciou nos Estados Unidos a elaboração de um catálogo para o acervo da *University of Brown*. Esse catálogo foi publicado em 1843 dividido em duas partes: “um catálogo descritivo de todas as obras que a Biblioteca possui e um índice de assuntos” (JEWETT, 1843, p. xix, tradução nossa). Na primeira parte (o “catálogo descritivo”), entradas contendo descrições completas das obras estavam organizadas, de modo geral, pelos nomes dos autores e em ordem alfabética, enquanto que na segunda parte entradas contendo descrições abreviadas com referências às entradas da primeira parte estavam organizadas por assunto também em ordem alfabética (JEWETT, 1843; BIAGETTI, 2008, p. 135). Nas duas partes é observada a presença de remissivas “Ver” (*See*) e “Ver também” (*See also*), por exemplo:

- “FOXTON, THOMAS. See BURNET, THOMAS”
- “Holy War. See Crusades”
- “JEBB, JOHN, D. D., *B’p of Limerick*. Correspondence with Knox. (See Forster Charles) *Note*. See also BURNET, GILBERT.”
- “LONDON AND WESTMINSTER REVIEW [...] (See also Westminster Review.)”
- “Agriculture [...] See also Gardening” (BROWN UNIVERSITY, 1843, p. 169, 517, 236, 266, 491).

As remissivas de nome para nome utilizadas por Panizzi, junto das remissivas utilizadas por Jewett, são consideradas aqui um dos primeiros meios de se alcançar o que hoje se entende por controle de autoridade, ainda que não seja identificada naquela época a presença de registros de autoridade.

Segundo Santos e Pereira (2014, p. 26), o catálogo de Jewett “[...] foi elogiado como um avanço na prática catalográfica”. Para Biagetti (2008, p. 135, tradução nossa), “[...] o elemento inovador foi o fato de que os cabeçalhos eram apresentados em uma única ordem alfabética, com uma relevante presença de remissivas”.

Em 1852, enquanto atuava no *Smithsonian Institution*, Jewett publicou o relatório *On the Construction of Catalogues of Libraries, and of a General Catalogue; and their Publication by Means of Separate, Stereotyped Titles: with Rules and Examples*¹¹ contendo “[...] regras de catalogação baseadas nas 91 regras de Panizzi [...] apresentava agradecimentos a Panizzi, e apontava algumas soluções diferentes para muitos pontos de descrição” (SANTOS; PEREIRA, 2014, p. 27). Em sua obra, Jewett apresentou também seu projeto para a reprodução de entradas utilizando a estereotipia, projeto esse que não foi concretizado (TAYLOR; JOUDREY, 2009, p. 75).

Ainda no século XIX, o desenvolvimento da Catalogação contou com a colaboração de Charles Ammi Cutter que, em 1861, começou a trabalhar com Ezra Abbot na *Harvard College library* na construção de “um catálogo público em cartões, planejado para ajudar os leitores no uso da biblioteca” (SANTOS; PEREIRA, 2014, p. 29).

Diferente da maioria dos catálogos de biblioteca contemporâneos que eram na forma de volumes publicados, Abbott e Cutter utilizaram cartões de fichários [*index cards*, do original em inglês], permitindo, assim, a flexibilidade para acrescentar ou remover itens à vontade, ao invés de ter de esperar até a próxima edição do catálogo. (STROMGREN, 2004, tradução nossa).

Para Santos e Pereira (2014, p. 29), “o catálogo em cartões de duas polegadas, escritos à mão e armazenados em gavetas etiquetadas, provou ser uma maneira nova e facilitadora para o uso público.” Segundo as autoras, em 1868 Cutter assumiu a posição de bibliotecário do *Boston Athenaeum*, onde trabalhou por 24 anos. Em sua primeira tarefa, a publicação de um catálogo da coleção, Cutter tornou-se pioneiro no uso de

¹¹ JEWETT, C. C. **On the construction of catalogues of libraries, and of a general catalogue; and their publication by means of separate, stereotyped titles:** with rules and examples. Washington: Smithsonian Institution, 1852. Disponível em: <<http://hdl.handle.net/2027/mdp.39015023260733>>. Acesso em: 12 dez. 2017.

cabeçalhos de assunto. O catálogo foi publicado em cinco volumes entre 1874 e 1882, tendo servido como modelo de catálogo dicionário em forma de livro (SANTOS; PEREIRA, 2014, p. 29).

Além de criar um sistema de classificação (*Expansive Classification*) e um sistema para a notação de autores (*Three-Figure Author Table*), Cutter contribuiu com a Catalogação também com seu conjunto de regras publicado em 1876 sob o título *Rules for a Printed Dictionary Catalogue* e que teve sua quarta e última edição em 1904 sob o título *Rules for a Dictionary Catalog* (CUTTER, 1904).

De notável contribuição,

Apesar do título, as regras não tratam somente de catálogos dicionários, configura-se como um código com 369 regras, de fácil leitura, incluindo uma parte descritiva e regras tanto para pontos de acesso de autor, título e assunto, quanto para alfabetação, catalogação de materiais especiais, normas de transliteração e elaboração de catálogos auxiliares. (SANTOS; PEREIRA, 2014, p. 29).

Cutter também é considerado o primeiro autor a definir as funções dos catálogos. Segundo ele, um catálogo deve:

1. Possibilitar que uma pessoa encontre um livro do qual

(A) o autor	}	é conhecido.
(B) o título		
(C) o assunto		
2. Apresentar o que a biblioteca possui
 - (D) de um determinado autor
 - (E) sobre um determinado assunto
 - (F) em um determinado tipo de literatura.
3. Auxiliar na escolha de um livro
 - (G) quanto a sua edição (bibliograficamente).
 - (H) quanto ao seu caráter (literário ou temático). (CUTTER, 1904, p. 12, tradução nossa).

Para o alcance desses objetivos, Cutter listou seis meios:

1. Entrada por autor com as remissivas necessárias (para A e D).
2. Entrada por título ou remissiva de título (para B).
3. Entrada por assunto, remissivas e tabela de classes de assuntos (para C e E).
4. Entrada de forma e entrada de idioma (para F).
5. Informando a edição e a imprensa, com notas quando necessário (para G).
6. Notas (para H). (CUTTER, 1904, p. 12, tradução nossa).

Sobre as contribuições de Panizzi e de Cutter para o desenvolvimento dos catálogos, Tillet (1989, p. 154, tradução nossa) destaca que eles

[...] foram influentes no estabelecimento da estrutura básica do catálogo enquanto um conjunto de registros com vínculos entre si. Esses registros incluíam entradas e remissivas; seus vínculos eram remissivas para nomes e obras. Mas, sua representação da estrutura do catálogo foi alterada na mudança de século quando uma nova tecnologia para a produção de catálogos foi introduzida.

A tecnologia a qual a autora refere-se é a ficha catalográfica.

Com o crescimento dos acervos das bibliotecas surgia a necessidade de inclusão de novas entradas nos catálogos. Nos catálogos em livros, as novas entradas eram acrescentadas nos espaços deixados entre as entradas existentes, preservando, assim, a ordenação já estabelecida. No entanto, conforme esses espaços eram ocupados, novas folhas eram intercaladas no catálogo ou as novas entradas eram acrescentadas fora da ordenação, o que causava dificuldades no uso do catálogo (GUERRINI, 2009, p. 1; SVENONIUS, 2000, p. 63). Uma alternativa seria a impressão de uma nova edição do catálogo incluindo as novas entradas, no entanto, além de custosa, tal alternativa representaria uma demora na atualização do catálogo.

Nos catálogos em fichas, as entradas eram registradas individualmente¹² em papéis retangulares (fichas) (SVENONIUS, 2000, p. 63), que eram ordenadas e, então, armazenadas em gavetas. Uma vez que as fichas não estavam presas umas as outras, como ocorria com as folhas de um catálogo em livro, elas poderiam ser movimentadas para permitir a inclusão e a remoção de entradas sem prejuízos à ordenação do catálogo. Com isso, os catálogos em fichas, por sua economia e dinamicidade (FRÍAS; PÉREZ ALVAREZ, 1995, p. 283), passaram a substituir gradualmente os catálogos em livros, de produção custosa e que rapidamente tornavam-se desatualizados (TAYLOR, 2006, p. 9).

Ainda que remontem ao século XVIII, quando Abbeé Rosier utilizou fichas na catalogação das publicações da *Académie des Sciences*, na França, em 1775 (SAYERS; STEWART, 1913, p. 8), os catálogos em fichas popularizaram-se entre as bibliotecas a partir do final do século XIX (CHAN, 2007, p. 22-23). O uso de catálogos em fichas como substitutos dos catálogos em livros também se deu no Brasil, como descreve Bettencourt (2014) em seu estudo sobre a representação da informação na Biblioteca Nacional.

¹² Cabe aqui uma menção à nota feita por Svenonius (2000, p. 208, tradução nossa): “Às vezes, várias fichas eram utilizadas na descrição de um documento e, ocasionalmente, fichas com entradas tracejadas eram utilizadas para descrever vários documentos”.

Segundo a autora, catálogos em livros eram utilizados para a representação de coleções de recursos informacionais na Biblioteca Nacional e há evidências do uso de catálogos em fichas a partir de 1876, na época denominados catálogos em cartões (BETTENCOURT, 2014, p. 110).

As primeiras fichas manuscritas eram denominadas “cartões alfabéticos” estavam acompanhadas das respectivas “referências e remissões”, todas ordenadas em um “catálogo móvel”, disponível para os usuários na chamada sala de consulta pública. (BETTENCOURT, 2014, p. 111, destaque do autor).

Além das vantagens mencionadas anteriormente, um dos fatores que contribuíram para a popularização dos catálogos em fichas foi o início do serviço de distribuição de fichas catalográficas em janeiro de 1901 pela *Library of Congress*, nos Estados Unidos (BARBOSA, 1978, p. 72; SVENONIUS, 2000, p. 63; TILLET, 1989, p. 154).

A utilização de fichas, por facilitar a reprodução das entradas, alterou o modo com que os componentes dos catálogos eram dispostos e apresentados aos usuários e alavancou o surgimento de outros componentes, tais como as entradas secundárias (TILLET, 1989, p. 154), que possibilitaram o oferecimento de uma quantidade maior de pontos de acesso aos usuários.

Em um catálogo em livro, de modo geral, a descrição completa de um recurso informacional era apresentada apenas na entrada que iniciava com o ponto de acesso autorizado do responsável principal pelo recurso. Ao buscar pelo nome de um responsável que não fosse o responsável principal, por exemplo, o segundo autor, o usuário poderia encontrar uma descrição abreviada do recurso sob o nome do responsável buscado; essa descrição abreviada funcionava como uma remissiva, pois direcionava o usuário para a entrada na qual a descrição completa do recurso informacional estava presente. Situações semelhantes poderiam ocorrer também com partes de obras, como mencionado anteriormente nos tipos de remissivas definidos nas 91 regras de Panizzi, e com assuntos, como notado na segunda parte do catálogo de Jewett.

Nos catálogos em fichas, a ficha que iniciava com o ponto de acesso autorizado para o responsável principal era denominada entrada principal¹³ (*main entry*), ficha

¹³ Foi utilizada aqui a definição de entrada principal apresentada no Glossário do AACR2r: “**Entrada principal.** *Main entry.* Registro catalográfico completo de um item, apresentado na forma que permita seja ele identificado e citado de maneira uniforme. A entrada principal pode incluir a pista (q.v.)” (CÓDIGO..., 2004, p. D-5, grifo do autor).

principal ou ficha matriz. Para prover pontos de acesso adicionais referentes aos demais responsáveis, aos títulos e aos assuntos do recurso informacional, a entrada principal era reproduzida e, no topo de cada cópia, era acrescentado o cabeçalho¹⁴ referente a um responsável, a um título ou a um assunto. As entradas criadas a partir da reprodução da entrada principal passaram a ser denominadas entradas secundárias (*added entries*). Com o surgimento das entradas secundárias, as remissivas contendo descrições abreviadas que apontavam para a descrição completa deixaram de ser utilizadas; as remissivas passaram a ser utilizadas, então, apenas para indicar ao usuário os cabeçalhos autorizados e os cabeçalhos relacionados (TILLET, 1989, 1992). Cabe notar, no entanto, que o uso de entradas secundárias teve início no século XX de forma gradual e com pouca clareza por parte dos códigos de catalogação, como indica a análise realizada por Tillett (1989, p. 154-156).

Os catálogos em fichas, por permitirem uma fácil intercalação das entradas, favoreceram a criação dos catálogos dicionários. Nesse tipo de catálogo, todas as entradas, independentemente do tipo do cabeçalho utilizado (responsáveis, títulos ou assuntos), eram reunidas e ordenadas em uma única sequência, de modo que os usuários do catálogo poderiam encontrar, em uma mesma gaveta, entradas com cabeçalhos representando os autores, os títulos e os assuntos dos recursos informacionais (GARRIDO ARILLA, 1996, p. 81; MEY; SILVEIRA, 2009, p. 188).

Na história da Biblioteca Nacional, Bettencourt (2014, p. 111) descreve o uso inicial de catálogos separados por tipos de pontos de acesso e, posteriormente, a reunião desses catálogos em um catálogo dicionário:

Quanto à sua tipologia, os catálogos em fichas inicialmente representavam os autores principais e os assuntos em módulos separados. A busca poderia ser feita no catálogo de autores ou no catálogo de assuntos. Em seguida, os catálogos passaram a ser enriquecidos com outros pontos de acesso, como título, coleção e todos esses pontos de acesso incluídos em um mesmo módulo, o chamado catálogo dicionário. (BETTENCOURT, 2014, p. 111).

Segundo Taylor e Joudrey (2009, p. 51), essa foi a forma padrão de organização das entradas dos catálogos na primeira metade do século XX, no entanto, com o crescimento dos acervos e, conseqüentemente, dos catálogos, ela tornou-se ineficiente, o

¹⁴ Foi utilizada aqui a definição de cabeçalho apresentada no Glossário do AACR2r: "**Cabeçalho**. *Heading*. Nome, palavra ou frase, colocados no alto de uma entrada catalográfica para fornecer um ponto de acesso." (CÓDIGO..., 2004, p. D-2, grifo do autor).

que fez com que as bibliotecas dividissem as entradas por tipo de cabeçalho e as mantivessem em catálogos diferentes, por exemplo, um catálogo com os títulos, outro com os assuntos, outro com os responsáveis (LUBETZKY, 1940; TAYLOR, 2006, p. 13, 16).

A partir dos catálogos em fichas é possível observar o surgimento dos registros e arquivos de autoridade. Sobre isso, Auld (1982, p. 320) aponta que Cutter reconheceu a necessidade do controle de autoridade por meio da criação de listas de autores do catalogador (*cataloger's author list*). Ainda que com o objetivo de evitar a duplicação de trabalho e não necessariamente de servir de apoio ao uso do catálogo, estas listas assemelhavam-se aos arquivos de autoridade atuais, contemplando elementos hoje presentes nos registros de autoridade:

A lista de autores do catalogador, mantida alfabeticamente, evita a duplicação de trabalho. Ela é um registro da forma do nome (na íntegra) adotada, com uma nota das fontes autorizadas consultadas e das variações do nome. (CUTTER, 1904, p. 133, tradução nossa).

Em 1908, a *American Library Association* (ALA) publicou a primeira edição de seu código de catalogação, *Catalog Rules: Author and Title Entries*¹⁵, que, influenciado pelas regras de Cutter, prescrevia o uso de remissivas, embora não descrevesse o uso de registros de autoridade ou de algo semelhante para manutenção dos cabeçalhos adotados e das remissivas criadas para eles (AULD, 1982, p. 320).

A segunda edição preliminar americana do código da ALA, publicada em 1941 com o título *A.L.A. Catalog Rules: Author and Title Entries*, incluiu um apêndice de três páginas sobre o uso e a criação de fichas de autoridade (*authority cards*) (AMERICAN LIBRARY ASSOCIATION, 1941, p. 339-331; AULD, 1982, p. 320). Nesse apêndice, observa-se a reiteração de elementos já listados por Cutter para inclusão nas listas de autores, e a inclusão de novos elementos que aproximam as fichas de autoridade desse código aos registros de autoridade existentes hoje. Alguns desses elementos são observados nesta instrução:

Faça uma ficha de autoridade para cada cabeçalho estabelecido informando o cabeçalho exatamente como ele aparece no catálogo. Liste as fontes consultadas na determinação do cabeçalho indicando aquelas nas quais datas foram encontradas e anotando quaisquer variações no nome ou nas datas encontradas em diferentes fontes. Informe de modo mais breve possível o autor, o título e a data do livro para o qual o

¹⁵ AMERICAN LIBRARY ASSOCIATION; LIBRARY ASSOCIATION. **Catalog rules**: author and title entries. American ed. Chicago: American Library Association Publishing board, 1908. Disponível em: <<http://hdl.handle.net/2027/hvd.32044021995329>>. Acesso em: 01 dez. 2016.

cabeçalho foi feito. Indique as remissivas. (AMERICAN LIBRARY ASSOCIATION, 1941, p. 339, tradução nossa).

Além dos nomes pessoais, o apêndice contemplava também exemplos de fichas de autoridade para nomes de entidades coletivas e títulos uniformes. A Figura 4 exemplifica uma ficha de autoridade criada de acordo com as instruções desse apêndice para um ponto de acesso de nome pessoal.

Figura 4 – Ficha de autoridade para um nome pessoal

<p>Lewis, Denslow, 1856-1913.</p> <p>[Full name: William Herbert Denslow Lewis] see x Lewis, William Herbert Denslow.</p> <p>✓Mich. U. Grad. (full name, d. 1913) ✓Amer. med. direct. 1909 (b. 1856)</p> <p>Jt. auth.: Anderson, T.R. Medical notes ... 1901.</p>

Fonte: *American Library Association* (1941, p. 339).

No cenário brasileiro, Bettencourt (2014, p. 111) destaca que na Biblioteca Nacional, “Com o objetivo de otimizar a padronização das entradas de autores e de assuntos, foram criados internamente, para uso dos bibliotecários, os catálogos de autoridades de nomes, de assuntos e de editores.” Ainda segundo a autora, após a consultoria realizada em 1945 por especialistas dos EUA,

Foram criados catálogos de autoridades de nomes e de assuntos para uso dos catalogadores e classificadores da instituição. O catálogo de autoridades de nomes passou a ser estruturado de acordo com o código de catalogação da ALA, constituindo-se em instrumento auxiliar para a padronização das entradas de nomes pessoais, entidades coletivas e eventos. O catálogo de assuntos seguia o modelo da lista de cabeçalhos de assuntos da *Library of Congress*. (BETTENCOURT, 2014, p. 117).

Sobre a primeira metade do século XX, Fiuza (1980, p. 143) destaca que nesse período

[...] os bibliotecários passaram a sofisticar cada vez mais o catálogo. Os catálogos de identidade se avolumaram, pois de cada nome de autor era feita uma pesquisa minuciosa nas fontes de referência; a catalogação descritiva passou a demandar decisões seríssimas e a quantidade de livros nos depósitos, esperando processamento técnico, se tornou assustadora.

A segunda edição do código da ALA, publicada em 1949, sob o título *A.L.A. Cataloging Rules for Author and Title Entries*¹⁶, retomou a abordagem da primeira edição de 1908: apresentava regras sobre remissivas, porém sem contemplar a criação de registros de autoridade. Essa abordagem foi seguida também na criação da primeira e da segunda edição do *Anglo-American Cataloguing Rules* (AACR), publicadas, respectivamente, em 1967 e 1978 (AULD, 1982, p. 321).

No contexto anglo-americano da catalogação, do qual derivam as práticas de catalogação adotadas hoje no Brasil, a falta de prescrição, por parte dos códigos de catalogação, levou à adoção de convenções para a criação de registros de autoridade nos catálogos em ficha, como destaca Burger (1985, p. 23-24, tradução nossa):

Não há nenhum método ou formato nacionalmente prescrito para a transcrição de dados de autoridade em fichas, mas várias convenções são amplamente utilizadas. A primeira e mais importante convenção é a localização do cabeçalho autorizado no topo da ficha, na mesma forma que ele aparece nas fichas catalográficas. Abaixo dele estão localizadas as remissivas. As remissivas Ver [*see from*] são indicadas pelo prefixo x e as remissivas Ver também [*see also from*] são indicadas pelo prefixo xx. Abaixo delas são dadas quaisquer notas. Finalmente, são dados o nome ou as iniciais do catalogador, a data em que registro foi preparado e o número de chamada ou o título da publicação que levou a preparação da ficha de autoridade.

Cabe aqui notar que as edições do código da ALA e do AACR não incluíam regras referentes aos pontos de acesso que representavam o conteúdo dos recursos informacionais, ou seja, os pontos de acesso de assunto; eram tratados apenas os pontos de acesso representando os títulos relacionados aos recursos (título principal, título uniforme, título da série, etc.) e os responsáveis por eles.

Para os pontos de acesso de assunto, outros instrumentos que não códigos de catalogação passaram a ser seguidos, tais como as listas de cabeçalhos de assunto e, posteriormente, os tesouros. Um exemplo dessas listas é *Library of Congress Subject Headings* (LCSH) iniciada em 1898 e publicada pela primeira vez como uma lista em 1909 (SVENONIUS, 2000, p. 178). Esses instrumentos, com outros como os sistemas de classificação bibliográfica, desenvolveram-se como parte do tratamento temático da informação, recebendo aportes de áreas como a Linguística e a Estatística (CINTRA et al., 1994, p. 33) e desempenhando, além das funções dos arquivos de autoridade de nomes e

¹⁶ AMERICAN LIBRARY ASSOCIATION. *A.L.A. cataloging rules for author and title entries*. Chicago: 1949. Disponível em: <<http://hdl.handle.net/2027/mdp.39015025988067>>. Acesso em: 25 fev. 2016.

títulos, outras funções, como a de “organizar o campo conceitual da representação documentária” (DODEBEI, 2014, p. 58). Hoje, esses instrumentos são encontrados na literatura sob as denominações vocabulários controlados, linguagens documentais e sistemas de organização do conhecimento.

Ainda que os sistemas de organização do conhecimento, principalmente os tesouros e as listas de cabeçalhos de assunto, desempenhem, em relação aos catálogos, funções semelhantes às aquelas desempenhadas pelos arquivos de autoridade de nomes, tais sistemas podem apresentar entre seus conceitos relacionamentos comumente não observados entre pessoas, entidades coletivas e demais entidades representadas nos arquivos de autoridade de nomes. Exemplos desses relacionamentos incluem as relações hierárquicas e as relações associativas.¹⁷

Mesmo com funções e aportes teóricos diferenciados, os arquivos de autoridade de assunto, quando comparados sob a ótica de seus relacionamentos com os catálogos e com os registros bibliográficos, mostram-se semelhantes aos arquivos de autoridade desenvolvidos para pessoas, entidades coletivas e títulos na Catalogação descritiva.

Os catálogos em fichas continuaram a ser utilizados durante o século XX e, ainda hoje, são mantidos em algumas bibliotecas, mesmo após o surgimento dos catálogos digitais, abordados na seção seguinte.

2.3 Controle de autoridade nos catálogos digitais

A partir da década de 1960, algumas bibliotecas, principalmente nos Estados Unidos, começaram a utilizar computadores na tentativa de suprir as demandas decorrentes do cenário informacional no qual se encontravam: o rápido desenvolvimento econômico, científico e tecnológico e o aumento da quantidade de publicações científicas e tecnológicas com a vida útil cada vez menor (RAYWARD, 2002). Barbosa (1978, p. 196) explica que

A necessidade de prover serviços em maior profundidade e de forma mais rápida a um maior número de usuários, bem como o aumento

¹⁷ Três tipos de relações são representadas nos sistemas de organização de conhecimento: de equivalência, de hierarquia e de associação. As relações de equivalência, por atuarem apenas no âmbito dos termos, são consideradas relações terminológicas. As relações hierárquicas e associativas, por atuarem no âmbito dos conceitos, são consideradas relações conceituais. As relações de equivalência estão presentes nos arquivos de autoridade de nomes por meio do uso de pontos de acesso não autorizados para pessoas, entidades coletivas, etc.

quantitativo dos materiais tradicionais, acrescentado ao aparecimento de novas formas de materiais, levaram as bibliotecas dos países desenvolvidos a optar pelo uso de computadores para processamento de suas operações internas. Por meio de processos simplesmente manuais tornava-se impossível garantir o tratamento técnico atualizado das coleções e o atendimento, em tempo hábil, aos usuários. Exigia-se portanto, um melhor nível de serviços.

Taylor (2006, p. 501, destaque do autor, tradução nossa) contextualiza o início da automação das bibliotecas no cenário tecnológico da década de 1970: “Quando a automação de bibliotecas teve início na década de 1970, a maioria dos sistemas estava baseada em um modelo hierárquico envolvendo um computador *mainframe* e um ou mais terminais ‘burros’”. Rayward (2002, p. 11, tradução nossa), ao traçar um histórico do uso de computadores em bibliotecas, sintetiza as vantagens inicialmente observadas que contribuíram para a automação das bibliotecas:

Os computadores pareciam oferecer aos bibliotecários a expectativa de um processamento mais eficiente, de serviços aprimorados aos usuários da biblioteca, de economia de dinheiro e contenção de custos, e de facilidade no compartilhamento de recursos e na cooperação entre bibliotecas, todas essas preocupações centrais ao tradicional funcionamento das bibliotecas.

A automação das bibliotecas no século XX pode ser dividida em quatro períodos: uso da computação *offline* (década de 1960 e início da década de 1970), surgimento dos sistemas *online* (década de 1970), surgimento dos microcomputadores e do CD-ROM (década de 1980) e revolução da Internet (década de 1990) (RAYWARD, 2002, p. 13). Com o surgimento dos sistemas online, no final da década de 1970, os catálogos digitais passaram a ser denominados *Online Public Access Catalogs* (OPACs). Taylor e Joudrey (2009, p. 165) destacam que os catálogos online ocorreram com a convergência da capacidade de processamento e de armazenamento dos computadores, do baixo custo, dos softwares para a manipulação de arquivos extensos e dos próprios arquivos em um formato digital.

Os OPACs desenvolveram-se em diferentes localidades, em diferentes períodos e com diferentes intensidades de acordo com os recursos tecnológicos disponíveis. Para Taylor e Joudrey (2009), os OPACs são classificados em apenas duas gerações. A primeira geração, que surgiu no início da década 1980, possuía interface de menus e geralmente possibilitava a busca apenas por autores e títulos, sem pontos de acesso de assunto e remissivas (TAYLOR; JOUDREY, 2009, p. 167). No final da década de 1980, a segunda geração de OPACs já apresentava melhorias em relação à geração anterior:

Essa geração foi marcada pelas interfaces de usuário significativamente aprimoradas. Foi introduzida a busca com o uso de operadores booleanos, aumentando, assim, a quantidade de pontos de acesso disponíveis para busca. Isto significou que não era mais necessário que as buscas fossem realizadas com a palavra ou a frase exata ou com uma âncora; as palavras agora poderiam corresponder à busca, mesmo se estivessem no meio de uma cadeia de caracteres. O processo de busca também foi melhorado com o uso de truncamento e de caracteres curingas, a capacidade de navegação (incluindo a navegação pelo índice de termos), o uso de todo o registro MARC, os refinamentos de busca interativos e o acesso por assunto. A segunda geração de OPACs trouxe também uma maior manipulação dos resultados de busca e melhores sistemas de ajuda com mensagens de erro mais informativas [...]. (TAYLOR; JOUDREY, 2009, p. 167-168, tradução nossa).

Sobre o uso de computadores especificamente nos processos de catalogação, McCallum (2002, p. 35) aponta que a ideia de obter eficiência nos fluxos de trabalho, principalmente pelo rápido compartilhamento de registros, foi a principal força que direcionou os experimentos para automação das bibliotecas. Mesmo com essas vantagens, os processos de catalogação, entre os processos realizados nas bibliotecas, não foram os primeiros a serem automatizados, cabendo essa posição aos processos de circulação e de inventário das coleções (MCCALLUM, 2002, p. 34). A autora explica que os processos de catalogação e de manutenção dos catálogos apresentavam complexos requisitos, o que fez com que eles não começassem a ser automatizados antes da metade da década de 1960, quando a *Library of Congress* iniciou os estudos para o MARC (*MAchine-Readable Cataloging*) *Pilot Project* (AVRAM, 1968; MCCALLUM, 2002, p. 34).

Como resultado de seus estudos, a LC deu início em 1966 a um serviço de distribuição de registros bibliográficos em fitas magnéticas, possibilitado pelo uso do Formato MARC (AVRAM, 1968). Esse formato consistia em especificações e designadores de conteúdo que, aplicados a registros bibliográficos, possibilitava a identificação e o processamento de seus dados por programas de computador, e, junto de uma estrutura genérica posteriormente reconhecida como as normas ANSI/NISO Z39.2 e ISO 2709, permitia a gravação desses registros bibliográficos em fitas magnéticas. Inicialmente, esse formato foi utilizado para a impressão de fichas catalográficas, de catálogos em livros e em microfichas e, posteriormente, para a criação de catálogos locais online (MCCALLUM, 2002, p. 43; TAYLOR, 2006, p. 498). A partir da experiência com o Formato MARC, a LC lançou em 1969 o Formato MARC II e, na década seguinte, já eram encontradas ao redor do planeta mais de 50 variações do formato original.

Para Clack (1988, p. 37), problemas no controle de autoridade são mais evidentes nos catálogos digitais que nos catálogos manuais. Segundo a autora, quando os computadores começaram a ser utilizados nas bibliotecas, houve o falso entendimento de que todos os problemas relacionados ao acesso aos registros bibliográficos seriam automaticamente solucionados e que a integridade desses registros não seria mais necessária, o que levou muitas bibliotecas a converterem seus catálogos manuais em digitais e rapidamente constatarem que seus catálogos digitais eram falhos em prover acesso completo e fácil aos registros bibliográficos e que o controle de autoridade era necessário em alguma medida para melhorar o acesso (CLACK, 1988, p. 37).

Assim como nos catálogos em livros e em fichas, nos catálogos digitais os registros bibliográficos estão sujeitos a alterações em decorrência de diversos fatores. Os pontos de acesso são alterados, por exemplo, devido às mudanças nas regras de catalogação, à descoberta de novos dados sobre as entidades (datas de nascimento e morte, sobrenomes omissos, etc. de uma pessoa, por exemplo) ou à substituição de um termo por outro mais apropriado para representar um determinado conceito. Para Malinconico (1975, p. 215, tradução nossa), a manutenção dos catálogos digitais para a realização de alterações como essas demanda a atenção a dois problemas básicos: “(1) a necessidade de garantir a consistência a qualquer momento, e (2) a necessidade de prover um método para a alteração rápida e precisa dessa estrutura de acordo com as circunstâncias”. Partindo dessas necessidades, o autor descreve o arquivo de autoridade digital implantado na *New York Public Library* em 1970 para auxiliar na criação, na gestão e no uso do catálogo dessa biblioteca, sendo esse, talvez, um dos primeiros arquivos de autoridade digitais relatados na literatura.

Partindo da necessidade do controle de autoridade em catálogos digitais e com o intuito de promover a automação dos arquivos de autoridade, em 1976, a LC publicou a edição preliminar do documento *Authorities: a MARC Format: Specifications for Magnetic Tapes Containing Authority Records*, provendo especificações e designadores de conteúdo para registros de autoridade de nomes, de títulos uniformes e de assuntos, de modo que pudessem ser gravados em fitas magnéticas utilizando a mesma estrutura genérica utilizada com os registros bibliográficos (LIBRARY OF CONGRESS, 1976; BURGER, 1985, p. 74).

A edição preliminar do formato foi utilizada para a distribuição de registros de autoridade nos Estados Unidos e na Europa até 1981, quando a LC lançou a primeira

edição do *Authorities: a MARC Format*¹⁸, que continha adendos à edição preliminar do formato, tais como a cobertura de pontos de acesso relativos a séries (BURGER, 1985, p. 74). Desde então a LC tem mantido e atualizado este formato, hoje denominado *MARC 21 Format for Authority Data* (Formato MARC 21 para Dados de Autoridade) e integrante do conjunto de Formatos MARC 21¹⁹. Acompanhando os desenvolvimentos realizados nos Estados Unidos em relação à automação dos arquivos de autoridade, outros formatos para dados de autoridade surgiram, tais como o *UNIMARC/Authorities: Universal Format for Authorities*, cuja primeira edição é de 1991.

A automação dos arquivos de autoridade trouxe diversas vantagens para as bibliotecas, como destaca Maxwell (2002, p. 11, tradução nossa):

O armazenamento de dados de autoridade em forma legível por máquina apresenta vantagens óbvias em relação aos arquivos em fichas, incluindo a facilidade de manipulação dos dados, a flexibilidade na apresentação e as capacidades de busca melhoradas, mas também abre a possibilidade de compartilhamento dos dados de autoridade entre bibliotecas, diminuindo, assim, o custo dessa importante tarefa.

Ao olhar para os catálogos digitais e para a função exercida pelos dados de autoridade é necessário considerar que, mesmo que estes catálogos tenham apresentado considerável desenvolvimento nas últimas décadas, dispondo de novas possibilidades de busca, de navegação e de interface, o modo com que os dados bibliográficos e de autoridade estão dispostos e relacionados sofreu pouca ou nenhuma alteração, visto que o modelo de dados, implantado com o Formato MARC II ainda na década de 1970 e posteriormente complementado com o formato para dados de autoridade, ainda é utilizado hoje na construção de catálogos digitais.

Para os propósitos desta pesquisa, um modelo é considerado “[...] uma descrição simplificada de certos aspectos da realidade utilizada para compreender, estruturar ou predizer partes do mundo real” (HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 2, tradução nossa). Para Coyle (2016, p. 29, tradução nossa),

¹⁸ LIBRARY OF CONGRESS. **Authorities, a MARC format**. Washington, 1981. Disponível em: <<http://hdl.handle.net/2027/umn.31951000485938n>>. Acesso em: 25 fev. 2016.

¹⁹ São coletivamente denominados Formatos MARC 21: *MARC 21 Format for Bibliographic Data* (Formato MARC 21 para Dados Bibliográficos), *MARC 21 Format for Authority Data* (Formato MARC 21 para Dados de Autoridade), *MARC 21 Format for Holdings Data* (Formato MARC 21 para Dados de Itens), *MARC 21 Format for Classification Data* (Formato MARC 21 para Dados de Classificação) e *MARC 21 Format for Community Information* (Formato MARC 21 para Informação Comunitária). Disponível em: <<https://www.loc.gov/marc/>>. Acesso em: 25 fev. 2016.

Há várias razões para se criar modelos do mundo real, a maior parte delas tem a ver com a dificuldade de manipular o mundo real diretamente. Arquitetos criam modelos de prédios que desenharam, projetistas criam modelos de argila de novos projetos automobilísticos e químicos criam modelos físicos para representar moléculas. Muitas vezes, nosso modelo do mundo não é um modelo físico, mas um modelo de dados simbólico. Esses modelos são abstrações do mundo real e sua semelhança com a realidade é mais conceitual do que física.

Sayão (2001, p. 86), em um estudo sobre os modelos teóricos na Ciência da Informação, destaca a proposta de Burt e Kinnucan (1990)²⁰ acerca de uma configuração contínua em que se situam os tipos de modelos de informação. De acordo com essa proposta, em um extremo do contínuo situa-se o ser humano com seu conhecimento, enquanto que no outro extremo está o sistema de informação e, entre os dois, o campo de representações: “[...] o espaço onde se encontram representações de uma ou de outra realidade. Essas representações experimentam criar pontes ou elucidar algumas ou todas as estranhezas que se supõe existam entre essas realidades” (SAYÃO, 2001, p. 86).

Os modelos que estão mais próximos do que seja a representação do usuário humano e do que se passa em sua cabeça em relação ao sistema são denominados modelos cognitivos; ao passo que os modelos que se identificam com o sistema e tentam descrever o que se passa em seu interior são mais conhecidos como modelos de dados. Na região intermediária do contínuo, é o lugar geométrico dos modelos que interpretam os usuários, o sistema e a interação entre eles; os modelos que se enquadram nesta categoria são coletivamente denominados de modelos conceituais. (SAYÃO, 2001, p. 86).

Mey e Silveira (2009, p. 17, destaque do autor), em sua introdução sobre o FRBR, esclarecem que modelo conceitual pode ser considerado “[...] como representação de algo; ‘conceitual’ implica a modelagem de coisas, processos ou abstrações, de forma a sintetizar e sistematizar sistemas teóricos ou fenômenos com vistas à aplicação”.

No âmbito dos bancos de dados, Fusco (2011, p. 63) considera um modelo de dados “[...] uma representação utilizada para descrever a estrutura conceitual, lógica e física de um banco de dados, esta estrutura compreende objetos de dados, características desses dados, relacionamentos e regras que restringem esses dados num cenário.” Para o autor, os modelos de dados são construídos em três níveis:

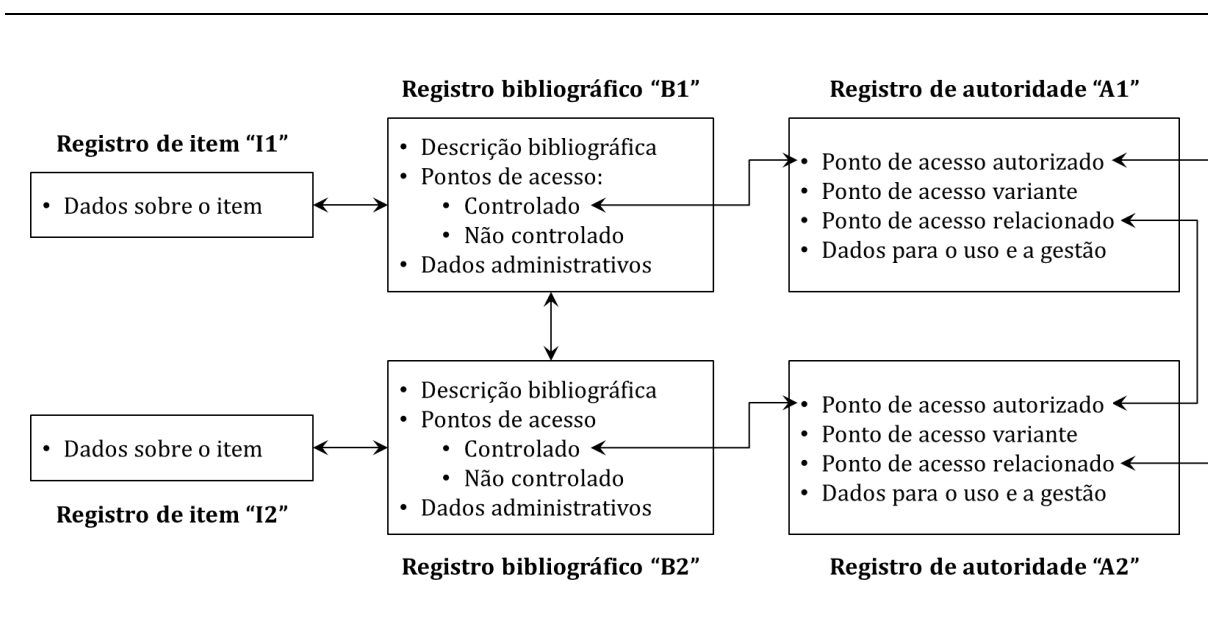
- Modelo de dados conceitual: fornece uma visão abstrata que representa os dados no domínio do cenário;

²⁰ BURT, P.; KINNUCAN, M. Information models and modeling techniques for information systems. **Annual Review of Information Science and Technology**, p. 175-208, 1990.

- Modelo de dados lógico: fornece uma visão numa representação de uma estrutura de modelo de banco de dados;
- Modelo de dados físico: fornece uma visão mais detalhada de como os dados estão armazenados no banco de dados. (FUSCO, 2011, p. 63).

Partindo desse entendimento, o modelo de dados dos catálogos digitais abordado neste estudo é um modelo de dados conceitual, uma vez que não se refere à estrutura do banco de dados em si (modelo de dados lógico) e nem ao armazenamento no banco de dados (modelo de dados físico), mas sim a uma visão abstrata acerca do domínio (o catálogo). Este modelo de dados dos catálogos digitais, representado na Figura 5, é aqui descrito da seguinte forma:

Figura 5 – Modelo de dados dos catálogos digitais



Fonte: Elaborada pelo autor.

- cada edição ou versão de um recurso informacional é representada por um registro bibliográfico;
- cada registro bibliográfico contém a descrição bibliográfica da edição ou versão do recurso informacional, os pontos de acesso estabelecidos para sua recuperação no catálogo e os dados administrativos sobre o recurso informacional e/ou sobre o registro;

- a cada registro bibliográfico são vinculados registros de itens para indicar os exemplares daquela edição ou versão do recurso informacional que fazem parte da coleção de uma determinada instituição;
- a cada registro bibliográfico são vinculados outros registros bibliográficos para indicar recursos informacionais relacionados;
- registros de autoridade podem ser criados para controlar os pontos de acesso utilizados nos registros bibliográficos;
- nos registros de autoridade, além do ponto de acesso autorizado, são incluídos pontos de acesso variantes, pontos de acesso relacionados e dados que apoiem o uso e a gestão do registro de autoridade;
- um ponto de acesso relacionado é um ponto de acesso autorizado controlado por outro registro de autoridade;
- o uso de pontos de acesso relacionados cria vínculos entre registros de autoridade;
- o uso de um ponto de acesso autorizado em um registro bibliográfico cria um vínculo, implícito ou explícito, entre o registro bibliográfico e o registro de autoridade que controla o ponto de acesso.

Os tipos de vínculos entre os registros bibliográficos e de autoridade decorrem do tipo de relacionamento entre o catálogo e o arquivo de autoridade. Para os propósitos desta pesquisa, considera-se que nos catálogos digitais são notados dois tipos de vínculos: implícitos e explícitos. Esses vínculos decorrem de dois tipos de relacionamentos entre arquivo de autoridade e catálogo listados por Burger (1985, p. 32-34), apresentados na Seção 2.1:

- **vínculos implícitos** ocorrem quando o arquivo de autoridade é independente do catálogo, quando os dois existem como bases de dados distintas, sem que as alterações realizadas no arquivo de autoridade sejam automaticamente refletidas no catálogo; vínculos implícitos são observados, por exemplo, nos casos em que os pontos de acesso autorizados utilizados em um catálogo são obtidos a partir de um arquivo de autoridade externo à instituição mantenedora do catálogo, sem que haja, localmente, um arquivo de autoridade; instituições que não possuem um arquivo de autoridade e utilizam o arquivo de autoridade da Biblioteca Nacional, por exemplo, enquadram-se nessa categoria;

- **vínculos explícitos** ocorrem quando o arquivo de autoridade está conectado ao catálogo, quando os dois estão vinculados de modo que uma alteração realizada no arquivo de autoridade é automaticamente refletida no catálogo; exemplos de vínculos explícitos são notados nos sistemas de gerenciamento de bibliotecas que permitem a criação de registros de autoridade e sua vinculação aos registros bibliográficos.

Além de influenciarem a gestão do catálogo, permitindo, por exemplo, a realização de alterações de forma automática, os tipos de vínculos entre registros bibliográficos e registros de autoridade podem influenciar também a consulta ao catálogo. Em relação a isso, são descritos na literatura dois cenários: (1) ao buscar por um ponto de acesso não autorizado, o usuário é orientado a realizar a busca utilizando o ponto de acesso autorizado; e (2) ao buscar por um ponto de acesso não autorizado, o usuário é automaticamente direcionado para os resultados de uma busca utilizando o ponto de acesso autorizado (BURGER, 1985, p. 43).

Já os tipos vínculos entre registros de autoridade dependem do tipo de entidade representada pelo registro. Por exemplo, registros de autoridade de pessoas podem estar vinculados por representarem uma pessoa e seu pseudônimo; registros de entidades coletivas podem estar vinculados por representarem uma entidade e sua sucessora ou a entidade hierarquicamente inferior ou superior a ela; registros de conceitos representados por termos tópicos podem estar vinculados por meio de relações hierárquicas (termo geral e termo específico) e associativas;²¹ etc.

Enquanto nos catálogos em livros e em fichas o acesso aos registros bibliográficos dava-se somente pelos pontos de acesso localizados no início das entradas, nos catálogos digitais qualquer dado presente em um registro, e não somente os nomes dos responsáveis, os títulos e os termos escolhidos, pode ser considerado um ponto de acesso (CARPENTER, 1989, p. 79; TAYLOR; JOUDREY, 2009, p. 39).

O acesso aos registros em um catálogo online é limitado somente ao que não está descrito. Qualquer elemento de metadado (tal como o publicador ou o ilustrador) utilizado em uma descrição bibliográfica pode servir como um ponto de acesso e a busca pode incorporar uma

²¹ No âmbito das linguagens documentais, os três tipos de relações destacadas pela literatura são de equivalência, hierárquica e associação. As relações de equivalência ocorrem no nível dos termos, vinculando termos não autorizados (ou não preferidos) aos termos autorizados. Nos catálogos digitais, os termos não autorizados são registrados no mesmo registro que os termos autorizados, de modo que as relações de equivalência não ocorrem entre registros de autoridade, mas sim entre partes de um mesmo registro.

variedade de sofisticados movimentos não lineares. (SVENONIUS, 2000, p. 23, tradução nossa).

Complementando a descrição dos catálogos digitais, Taylor (2006, p. 12, tradução nossa) explica que eles não possuem entradas no mesmo sentido que os catálogos em livros ou em fichas, em que a descrição de recurso era replicada, em sua totalidade ou não, a cada entrada necessária, uma vez que nos catálogos online

Uma cópia de cada registro bibliográfico é armazenada pelo sistema. Índices são gerados para vincular cada nome, título, cabeçalho de assunto ou outra entidade utilizada como ponto de acesso aos registros bibliográficos associados. Em resposta a uma busca, os elementos de dados selecionados de cada registro bibliográfico relevante são copiados para a tela do usuário em um formato de apresentação escolhido para aquele sistema.

O arranjo dos registros bibliográficos nos catálogos em livros permitia que as obras de um mesmo responsável, com suas diferentes edições ou versões, fossem apresentadas como um conjunto ao usuário, provendo a este uma visão geral das obras. O arranjo dos registros bibliográficos nos catálogos em fichas, em que as diferentes edições de uma obra eram registradas em fichas individuais apresentadas segundo uma ordem alfabética específica, reduziu a capacidade do catálogo de, com o máximo possível de economia, prover uma visão geral de todas as obras, edições e versões relacionadas a um responsável. Sobre o arranjo dos registros bibliográficos em catálogos digitais, Chan (2007, p. 25, tradução nossa) aponta que

O modo com que os registros são arranjados internamente em um catálogo é uma questão de design do sistema e afeta os usuários finais somente no sentido de que um sistema pode ser mais fácil de utilizar e aparentemente recuperar resultados mais satisfatórios que o outro. Mas, como os itens recuperados são apresentados na tela também depende do design do sistema e, aqui, os usuários finais devem estar cientes de que mais opções estão abertas do que aquelas disponíveis no uso de um catálogo manual. Os resultados de uma busca são normalmente apresentados em uma ordem padrão, por data ou por nome, mas, na maioria dos sistemas, a lista pode ser reordenada de acordo com as preferências dos usuários.

Convém ressaltar que nem sempre o controle de autoridade esteve presente nos catálogos digitais e que, ainda hoje, existem catálogos que, por motivos diversos e alheios aos objetivos desta pesquisa, carecem de qualquer atividade de controle dos pontos de acesso. Sobre a inclusão do controle de autoridade nos catálogos digitais, Brandt (1988, p. 297, tradução nossa), no final da década de 1980, escreveu que

À medida que mais e mais bibliotecas migram para catálogos online, o controle de autoridade está sendo visto de perto como uma parte do gerenciamento de dados para assegurar a consistência e a integridade das entradas, assim como para relacionar registros que compartilham um cabeçalho. Quando aplicado às bases de dados bibliográficos, o controle de autoridade é mais que um meio usado para estabelecer e relacionar os cabeçalhos – ele pode automaticamente e literalmente “vincular” os registros. Como tal, o controle de autoridade pode ser visto como uma ferramenta essencial para a administração de dados.

Assim como ocorreu com os catálogos digitais, melhorias foram implementadas com a evolução dos arquivos de autoridade digitais no decorrer dos anos. O estabelecimento de vínculos automáticos entre registros bibliográficos e de autoridade, descrita por Burger (1985, p. 32-34) e por Brandt (1988, p. 297), é um exemplo dessas melhorias. Em 1982, partindo dessa possibilidade de vínculo, Auld (1982, p. 326-327, tradução nossa) descreveu como o controle de autoridade poderia idealmente manifestar-se em um catálogo:

Como o controle de autoridade pode manifestar-se idealmente em um catálogo? As seguintes características parecem razoáveis do ponto de vista do catalogador e do usuário. Um registro bibliográfico, junto de todas as formas variantes de cada cabeçalho associado a ele, seria inserido no sistema. O computador estabeleceria vínculos [*linkages*] entre os registros de autoridade e as formas variantes dos cabeçalhos e entre as formas preferidas e o registro bibliográfico. Quando um usuário digitasse uma forma conhecida de um cabeçalho, o sistema seguiria os vínculos internos a exibiria o item requerido mesmo se a forma preferida do cabeçalho fosse completamente diferente da forma digitada. Embora o computador tivesse seguido um vínculo para a forma preferida do cabeçalho e, então, um segundo vínculo para o registro bibliográfico, para o usuário, aparentaria que um vínculo direto existe entre a forma do cabeçalho digitada e o registro bibliográfico exibido. O mecanismo de controle de autoridade seria invisível até que o usuário se interessasse por ele. Os registros de autoridade seriam exibidos somente quando requeridos. Cutter teria apreciado um catálogo assim!

O crescente uso de catálogos e de arquivos de autoridade digitais, junto do surgimento de redes e de programas para o compartilhamento de registros bibliográficos e de autoridade, impulsionou no domínio bibliográfico os esforços para a criação de padrões, diretrizes e modelos que favorecessem o compartilhamento de dados catalográficos em âmbito internacional. Alguns desses esforços proveem um entendimento atual e sistematizado, ainda que não seja consensual, acerca dos dados de autoridade, como é caso dos modelos conceituais FRAD e FRSAD, apresentados na seção seguinte.

2.4 Functional Requirements for Authority Data (FRAD) e Functional Requirements for Subject Authority Data (FRSAD)

O compartilhamento de dados, em qualquer domínio, requer padronização. A capacidade de compartilhamento dos dados resultantes dos processos de catalogação, compreendida como uma importante vantagem da automação de bibliotecas, pressupõe a existência de padronização entre as instituições envolvidas nesse compartilhamento.

No domínio bibliográfico, o anseio pelo compartilhamento de dados é observado já em períodos remotos como na Idade Média, quando, no século XIII, monges ingleses elaboraram um catálogo coletivo representando obras presentes em 183 bibliotecas monásticas (SANTOS; PEREIRA, 2014, p. 17). De forma mais intensa, o anseio por esse compartilhamento é notado a partir da segunda metade do século XIX com duas principais ações ainda voltadas aos catálogos analógicos: o desenvolvimento de códigos de catalogação de amplo alcance e a consolidação do uso de fichas catalográficas em um tamanho padrão (7,5 x 12,5 cm). Com a introdução do uso de computadores nas atividades de catalogação, a partir da segunda metade do século XX, o anseio pelo compartilhamento de dados levou ao desenvolvimento de meios para a padronização e a troca de informações já em ambientes digitais, entre esses meios, o Formato MARC e os formatos dele derivados, as normas ANSI/NISO Z39.2 e ISO 2709 e o protocolo Z39.50.

Como parte dos seus esforços para o controle bibliográfico universal, a *International Federation of Library Associations and Institutions* (IFLA) tem publicado padrões, orientações e formatos no intuito de, a partir da padronização, favorecer o compartilhamento de dados catalográficos em uma escala global. Algumas dessas publicações²² destinam-se aos dados de autoridade, entre elas:

- *Guidelines for authority and reference entries* (1984), com uma segunda edição publicada em 2001 sob o título *Guidelines for authority records and references*;
- *Management and use of name authority files: personal names, corporate bodies and uniform titles: evaluation and prospects* (1991);
- *UNIMARC/Authorities: Universal Format for Authorities* (1991), com segunda e terceira edições publicadas em 2001 e 2009, respectivamente;

²² Uma lista dessas publicações está disponível em: <<https://www.ifla.org/node/586>>. Acesso em: 12 dez. 2017.

- *Mandatory data elements for internationally shared resource authority records* (1998).

Ao final do século XX, o domínio bibliográfico, com seus instrumentos de representação estabelecidos principalmente a partir de uma realidade existente décadas antes, passou a enfrentar um novo cenário:

As bibliotecas na década de 1990 foram forçadas a começar a reconsiderar seus padrões e práticas relacionados à organização da informação bibliográfica, por um lado, considerando na prática cotidiana a maturidade do pensamento teórico e os padrões desenvolvidos internacionalmente, e, por outro, a pressão para incluir a descrição de novos formatos de documentos – recursos eletrônicos (nascidos digitais ou digitalizados) e, posteriormente, documentos da Web – em seus catálogos de modo a satisfazer as necessidades dos usuários. (WILLER; DUNSIRE, 2013, p. 2, tradução nossa).

Com isso, um processo de questionamento dos conceitos e das teorias tradicionais foi iniciado em 1990 com um seminário sobre registros bibliográficos realizado pela IFLA em Estocolmo, na Suécia (WILLER; DUNSIRE, 2013, p. 13). Willer e Dunsire (2013, p. 13, tradução nossa) explicam que

Foi um momento em que a necessidade de mudança estava no ar e impulsionada por muitos fatores, incluindo o desenvolvimento e o crescimento de sistemas automatizados e de bases de dados nacionais e internacionais de larga escala, a necessidade de reduzir os custos com a catalogação por meio do reuso dos registros bibliográficos existentes, o crescimento contínuo das publicações, especialmente das publicações eletrônicas, e uma ampliação da variedade de expectativas e de necessidades dos usuários.

e que os participantes do seminário

[...] admitiram que “o ambiente em que os princípios e padrões da catalogação operavam havia mudado dramaticamente” e “reconheceram a importância de satisfazer as necessidades do usuário e abordar mais efetivamente a ampla variedade de necessidades associadas aos vários tipos de materiais e aos vários contextos em que os registros bibliográficos são utilizados”. (WILLER; DUNSIRE, 2013, p. 13, destaque do autor, tradução nossa).

Assim, com o intuito de relacionar os dados dos registros bibliográficos às necessidades dos usuários e de recomendar um nível básico de funcionalidade para os registros criados por agências bibliográficas nacionais (IFLA STUDY GROUP ON THE FRBR, 1998, p. 7), um grupo da IFLA conduziu um estudo que resultou no modelo conceitual *Functional Requirements for Bibliographic Records* (Requisitos Funcionais para Registros Bibliográficos), mais conhecido como FRBR, publicado em 1998. Esse modelo apresenta

um conjunto de entidades de interesse do universo bibliográfico, os atributos que caracterizam cada uma dessas entidades e os relacionamentos existentes entre elas, assim como ações, denominadas tarefas dos usuários, que demandam tais atributos e relacionamentos.

Desde sua publicação, o FRBR tem influenciado de forma significativa o desenvolvimento da Catalogação. Segundo Coyle (2016, p. 65, tradução nossa),

O modelo de dados bibliográficos FRBR é a mudança de pensamento mais radical sobre os catálogos bibliográficos desde que Panizzi desenvolveu suas noventa e uma regras para o catálogo do *British Museum*. [...] O FRBR é alternadamente visto como uma análise das necessidades do usuário, como uma descrição do fluxo de catalogação e como um modelo de dados para um futuro formato de registro bibliográfico. No entanto, ele é raramente visto como originalmente foi projetado para ser: um desenvolvimento dos requisitos básicos para um padrão internacional de registros bibliográficos nacionais.

Entre as recomendações do grupo de estudo responsável pelo FRBR estava a necessidade de expansão do modelo para a cobertura dos dados de autoridade (IFLA STUDY GROUP ON THE FRBR, 1998, p. 5). Partindo dessa recomendação, em 1999 a IFLA iniciou o *Working Group on Functional Requirements and Numbering of Authority Records* (FRANAR) (Grupo de Trabalho sobre Requisitos Funcionais e Numeração de Registros de Autoridade), voltado aos registros de autoridade (IFLA WORKING GROUP ON FRANAR, 2013, p. i). O resultado do estudo realizado pelo FRANAR foi publicado em 2009 sob o título *Functional Requirements for Authority Data: a conceptual model* (Requisitos Funcionais para Dados de Autoridade: um modelo conceitual), passando a ser conhecido pela sigla FRAD (IFLA WORKING GROUP ON FRANAR, 2013).²³

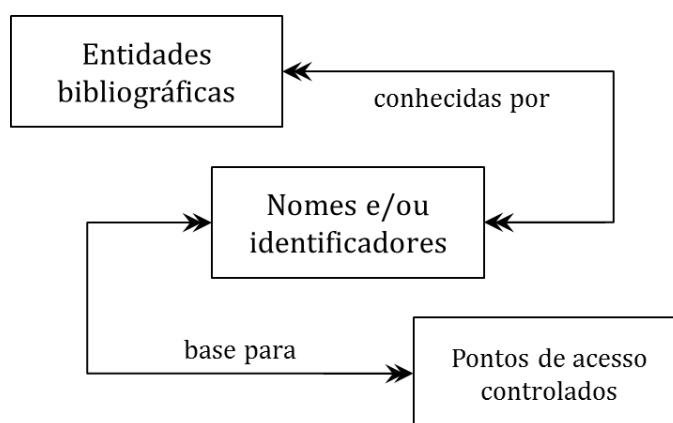
O FRAD, cujo escopo limita-se aos dados de autoridade referentes a nomes e títulos, excluindo, portanto, os dados de autoridade referentes aos conceitos (“assuntos”), tem como propósito principal

[...] prover um quadro para a análise dos requisitos funcionais para o tipo de dado de autoridade requerido para apoiar o controle de autoridade e para o compartilhamento internacional de dados de autoridade. O modelo foca nos dados, independentemente de como eles estão acondicionados (e.g., em registros de autoridade). (IFLA WORKING GROUP ON FRANAR, 2013, p. 1, tradução nossa).

²³ Embora o grupo de trabalho tenha em seu nome “registro de autoridade”, o modelo publicado refere-se aos “dados de autoridade”, independentemente de como estão acondicionados (IFLA WORKING GROUP ON FRANAR, 2013, p. 1, tradução nossa). Um rascunho do modelo, publicado em 2005, foi intitulado *Functional Requirements for Authority Records*.

A partir da base fundamental representada na Figura 6 e utilizando-se da técnica de análise de entidades, atributos e relacionamentos, técnica essa surgida da Ciência da Computação para a criação de modelos conceituais para bases de dados relacionais (COYLE, 2016, p. 33), são definidos no FRAD entidades, atributos, relacionamentos e tarefas dos usuários relativos aos dados de autoridade.

Figura 6 – Base fundamental do FRAD



Fonte: *IFLA Working Group on FRANAR* (2013, p. 4, tradução nossa).

A base fundamental apresentada na Figura 6 é entendida da seguinte forma:

As entidades no universo bibliográfico (tais como aquelas identificadas no *Functional Requirements for Bibliographic Records*) são conhecidas por nomes e/ou identificadores. No processo de catalogação (seja em bibliotecas, museus ou arquivos), estes nomes e identificadores são utilizados como base para a construção de pontos de acesso controlados. (IFLA WORKING GROUP ON FRANAR, 2013, p. 3, tradução nossa).

As entidades e os atributos definidos no FRAD a partir dessa base conceitual são listados no Quadro 1.

Os relacionamentos entre as entidades apresentadas no Quadro 1 são apresentados na Figura 7.

Quadro 1 – Entidades e atributos definidos no FRAD

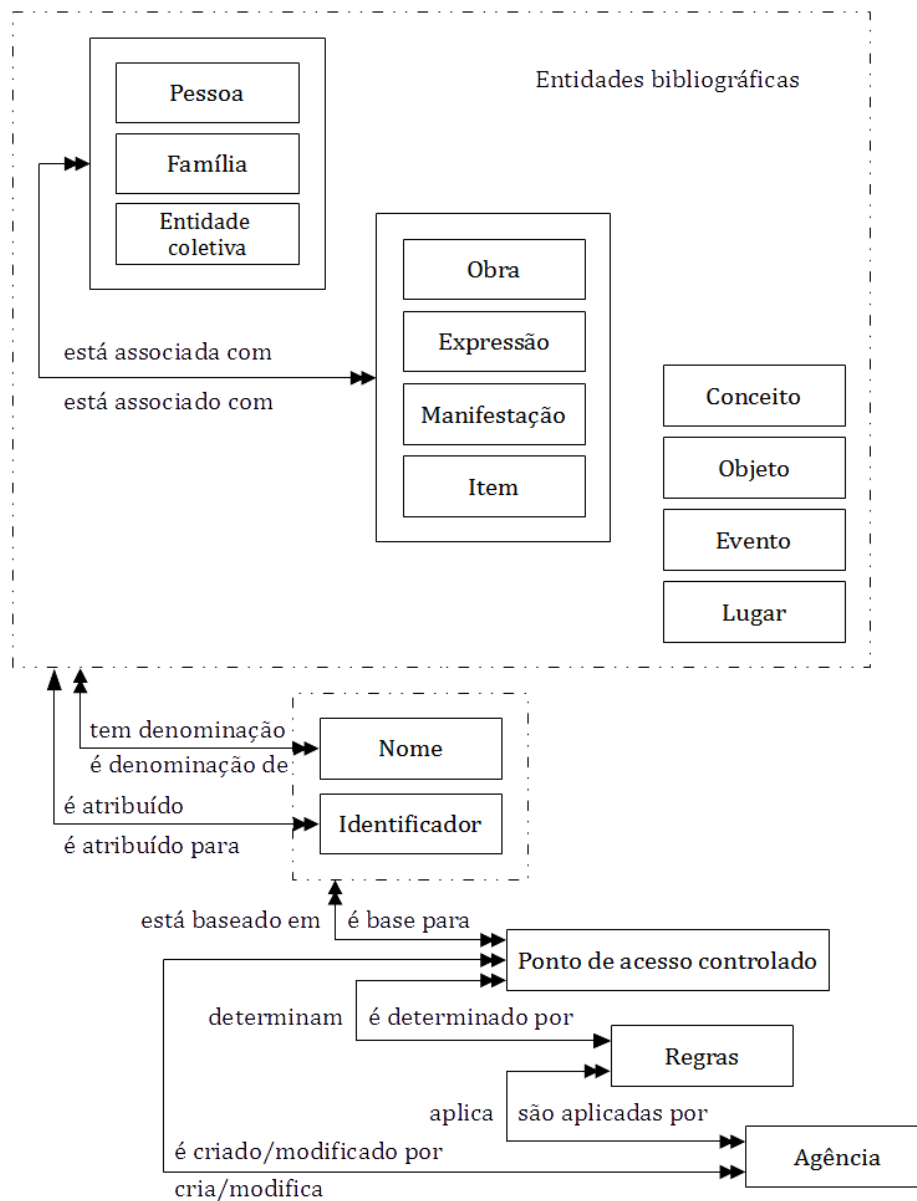
Entidades	Atributos
<p>Pessoa (<i>person</i>)</p> <p>Um indivíduo, uma personalidade ou uma identidade criada ou adotada por um indivíduo ou grupo.</p>	<p>Datas associadas à pessoa Título da pessoa Gênero Local de nascimento Local de morte País Local de residência Afiliação Endereço Idioma da pessoa Campo de atividade Profissão/ocupação Biografia/história Outra informação associada à pessoa</p>
<p>Família (<i>family</i>)</p> <p>Duas ou mais pessoas relacionadas por nascimento, casamento, adoção, união civil ou status legal similar, ou que se apresentem como uma família.</p>	<p>Tipo de família Datas da família Locais associados à família Idioma da família Campo de atividade História da família</p>
<p>Entidade coletiva (<i>corporate body</i>)</p> <p>Uma organização ou grupo de pessoas e/ou organizações identificadas por um determinado nome que atuam como uma unidade.</p>	<p>Local associado à entidade coletiva Datas associadas à entidade coletiva Idioma da entidade coletiva Endereço Campo de atividade História Outra informação associada à entidade coletiva</p>
<p>Obra (<i>work</i>)</p> <p>Uma criação intelectual ou artística distinta (i.e., conteúdo intelectual ou artístico).</p>	<p>Forma da obra Data da obra Meio de execução Assunto da obra Designação numérica Clave Local de origem da obra História Outra característica distinta</p>
<p>Expressão (<i>expression</i>)</p> <p>A realização intelectual ou artística de uma obra na forma alfanumérica, musical, notação coreográfica, som, imagem, objeto, movimento, etc., ou qualquer combinação de tais formas.</p>	<p>Forma da expressão Data da expressão Meio de execução da expressão Idioma da expressão Técnica Outra característica distinta</p>
<p>Manifestação (<i>manifestation</i>)</p> <p>A materialização de uma expressão de uma obra.</p>	<p>Designação de edição/número Local de publicação/distribuição Publicador/distribuidor Data de publicação/distribuição Forma do suporte Numeração</p>

Entidades	Atributos
<p>Item (<i>item</i>) Um único exemplar de uma manifestação.</p>	<p>Localização do item Histórico de guarda do item Fonte de aquisição imediata do item</p>
<p>Conceito (<i>concept</i>) Uma noção abstrata ou ideia.</p>	<p>Uma vez que essas entidades relacionam-se aos dados de autoridade de assunto, o FRANAR optou por deixar a definição de seus atributos fora do escopo do FRAD.</p>
<p>Objeto (<i>object</i>) Uma coisa material.</p>	
<p>Evento (<i>event</i>) Uma ação ou ocorrência.</p>	
<p>Lugar (<i>place</i>) Um local.</p>	
<p>Nome (<i>name</i>) Um caractere ou grupo de palavras e/ou caracteres pelo qual uma entidade é conhecida.</p>	<p>Tipo de nome Nome [<i>name string</i>] Escopo de uso Data de uso Idioma do nome Alfabeto do nome Esquema de transliteração do nome</p>
<p>Identificador (<i>identifier</i>) Um número, código, palavra, frase, etc. que é unicamente associado a uma entidade, e serve para diferenciar essa entidade de outras no domínio em que o identificador é atribuído.</p>	<p>Tipo de identificador</p>
<p>Ponto de acesso controlado (<i>controlled access point</i>) Um nome, termo, código, etc. sob o qual um registro bibliográfico, um registro de autoridade ou uma remissiva será encontrado.</p>	<p>Tipo de ponto de acesso controlado Status do ponto de acesso controlado Uso pretendido do ponto de acesso controlado Ponto de acesso não diferenciável Idioma do ponto de acesso base Idioma da catalogação Alfabeto do ponto de acesso base Alfabeto da catalogação Esquema de transliteração do ponto de acesso base Esquema de transliteração da catalogação Fonte de ponto de acesso controlado Ponto de acesso base Acréscimo</p>
<p>Regras (<i>rules</i>) Um conjunto de instruções relacionadas à formulação e/ou ao registro de pontos de acesso controlados (formas autorizadas, formas variantes ou remissivas, etc.).</p>	<p>Citação das regras Identificador das regras</p>

Entidades	Atributos
<p>Agência (<i>agency</i>)</p> <p>Uma organização responsável pela criação ou modificação de um ponto de acesso controlado. A agência é responsável pela aplicação e interpretação das regras que ela cria e/ou utiliza. A agência também pode ser responsável pela criação e manutenção de identificadores dentro de seu domínio.</p>	<p>Nome da agência Identificador da agência Localização da agência</p>

Fonte: Elaborado pelo autor a partir do FRAD (IFLA WORKING GROUP ON FRANAR, 2013, p. 8-15, 17-29).

Figura 7 – Relacionamentos entre as entidades do FRAD



Fonte: IFLA Working Group on FRANAR (2013, p. 7, tradução nossa).

No âmbito do FRAD, são compreendidos como usuários dos dados de autoridade os criadores desses dados e os usuários que os utilizam diretamente ou indiretamente por meio dos pontos de acesso controlados nos catálogos, bibliografias nacionais e outras bases de dados similares (IFLA WORKING GROUP ON FRANAR, 2013, p. 46). O modelo conceitual apresenta também quatro tarefas que esses usuários realizam a partir dos dados de autoridade:

- **encontrar** (*find*): encontrar uma entidade ou um conjunto de entidades correspondente ao critério estipulado (ou seja, encontrar uma única entidade ou um conjunto de entidades utilizando como critério de busca um atributo, a combinação de atributos ou um relacionamento da entidade); ou explorar o universo das entidades bibliográficas utilizando seus atributos e relacionamentos;
- **identificar** (*identify*): identificar uma entidade (ou seja, confirmar se a entidade representada corresponde à entidade buscada, distinguir entre duas ou mais entidades com características similares) ou validar a forma do nome para ser utilizada como um ponto de acesso controlado;
- **contextualizar** (*contextualize*): localizar uma pessoa, entidade coletiva, obra, etc. no contexto; esclarecer o relacionamento entre duas ou mais pessoas, entidades coletivas, obras, etc.; ou esclarecer o relacionamento entre uma pessoa, entidade coletiva, etc. e o nome pelo qual essa pessoa, entidade coletiva, etc. é conhecida (por exemplo, o nome utilizado na religião versus o nome secular);
- **justificar** (*justify*): documentar a razão pela qual o criador dos dados de autoridade escolheu o nome ou a forma do nome na qual o ponto de acesso controlado está baseado (IFLA WORKING GROUP ON FRANAR, 2013, p. 46, tradução nossa).

Enquanto o FRANAR trabalhava no desenvolvimento do FRAD, outro grupo de trabalho foi estabelecido pela IFLA, desta vez com o foco nos dados de autoridade de assunto. O *Working Group on the Functional Requirements for Subject Authority Records* (FRSAR), formado em 2005, atuou até 2010, quando foi publicado seu relatório final intitulado *Functional Requirements for Subject Authority Data* (Requisitos Funcionais para Dados de Autoridade de Assunto) (FRSAD).

Neste relatório, o grupo de trabalho apresenta um modelo construído a partir do FRBR e utilizando também a técnica de análise de entidades, atributos e

relacionamentos. O FRASAD define as duas entidades, apresentadas no Quadro 2 com seus atributos, relacionadas aos dados de autoridade de assunto.

Quadro 2 – Entidades e atributos definidos no FRASAD

Entidades	Atributos²⁴
<p><i>Thema</i> Qualquer entidade utilizada como um assunto de uma obra.</p>	<p>Tipo de <i>thema</i> Nota de escopo</p>
<p><i>Nomen</i> Qualquer signo ou sequência de signos (caracteres alfanuméricos, símbolos, sons, etc.) pelo qual um <i>thema</i> é conhecido, referenciado ou tratado.</p>	<p>Tipo de <i>nomen</i> Esquema Fonte de referência do <i>nomen</i> Representação do <i>nomen</i> Idioma do <i>nomen</i> Alfabeto do <i>nomen</i> Alfabeto de conversão Forma do <i>nomen</i> Período de validade do <i>nomen</i> Público Status do <i>nomen</i></p>

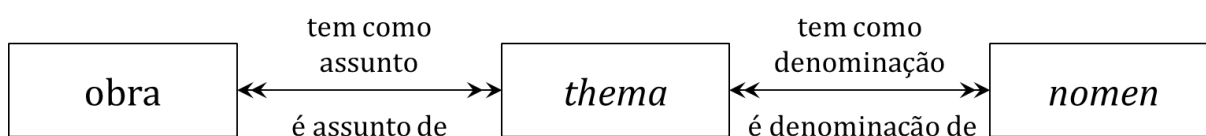
Fonte: Elaborado pelo autor a partir do FRASAD (IFLA WORKING GROUP ON FRASAR, 2010).

Apesar de o FRASAD definir apenas as entidades *thema* e *nomen*, esse modelo conceitual relaciona tais entidades à entidade obra, definida no FRBR. Os relacionamentos entre as três entidades, representados na Figura 8, são compreendidos da seguinte forma:

- uma obra tem como assunto um ou mais *themas*;
- um *thema* é assunto de uma ou mais obras;
- um *thema* tem como denominação um ou mais *nomens*;
- um *nomen* é a denominação de um ou mais *themas*.

²⁴ Convém mencionar as observações sobre os atributos definidos no FRASAD: “Em qualquer implementação haverá normalmente atributos adicionais de um *thema* além de ‘tipo’ e ‘nota de escopo’. Esses atributos dependerão do tipo de *thema* e do domínio de aplicação.” (IFLA WORKING GROUP ON FRASAR, 2010, p. 19, destaque do autor, tradução nossa) e “Os atributos de um *nomen* representam os atributos mais comuns (gerais). Nem todos são aplicáveis em todos os casos e a lista não é abrangente. [...] Além dos atributos gerais propostos aqui, pode haver atributos adicionais específicos da implementação.” (IFLA WORKING GROUP ON FRASAR, 2010, p. 21, tradução nossa).

Figura 8 – Relacionamento entre as entidades obra, *thema* e *nomen*



Fonte: IFLA Working Group on FRASAR (2010, p. 15, tradução nossa).

Cabe notar que a entidade *thema* é considerada uma super-entidade ou uma superclasse que compreende todas as entidades definidas no FRBR (as quais estão definidas também no FRAD): obra, expressão, manifestação, item, pessoa, família, entidade coletiva, conceito, objeto, evento e lugar, uma vez que essas entidades podem ser assuntos da entidade obra (IFLA WORKING GROUP ON FRASAR, 2010, 17).

Como potenciais usuários dos dados de autoridade de assunto, o FRASAD (IFLA WORKING GROUP ON FRASAR, 2010, p. 33, tradução nossa) define: (a) profissionais da informação que criam e mantêm dados de autoridade de assunto, incluindo catalogadores e criadores de vocabulários controlados; (b) profissionais da informação que criam e mantêm metadados; (c) bibliotecários de referência e outros profissionais da informação que buscam informações como intermediários; e (d) usuários finais que buscam informações para satisfazer suas necessidades informacionais.

Esses usuários utilizam os dados de autoridade de assunto para as tarefas de:

- **encontrar** (*find*) um ou mais assuntos e/ou suas denominações que correspondam ao critério de busca estipulado, utilizando atributos e relacionamentos;
- **identificar** (*identify*) um assunto e/ou sua denominação baseado em seus atributos ou relacionamentos (ou seja, distinguir entre dois ou mais assuntos ou denominações com características similares e confirmar que o assunto ou a denominação apropriada foi encontrada);
- **selecionar** (*select*) um assunto e/ou uma denominação apropriada às necessidades do usuário (ou seja, escolher ou rejeitar com base nos requisitos e nas necessidades do usuário);
- **explorar** (*explore*) os relacionamentos entre assuntos e/ou suas denominações (ou seja, explorar os relacionamentos para entender a estrutura de um domínio

de assunto e sua terminologia) (IFLA WORKING GROUP ON FRASAR, 2010, p. 33, tradução nossa).

Como apresentado neste capítulo, a presença dos dados de autoridade é observada ao longo dos últimos 150 anos da história da Catalogação, evoluindo desde o uso de remissivas nos catálogos em livros, passando pelas fichas de autoridade e chegando até os sistemas de autoridade em ambientes digitais que permitem o compartilhamento de dados em uma escala global.

Ao longo da história da Catalogação, os dados de autoridade acumularam funções como a de permitir aos usuários, sejam eles seus criadores ou seus utilizadores, encontrar, identificar, selecionar, contextualizar, justificar e explorar, como definido nos modelos conceituais FRAD e FRASAD. Além dessa função, os dados de autoridade, em especial os pontos de acesso, atuam como dispositivos de ligação – “*linking devices*” nas palavras de Tillett (1992, p. 23) – conectando os registros bibliográficos de recursos informacionais relacionados.

A história da Catalogação é permeada pelo uso de tecnologias: técnicas de impressão, fichas catalográficas e formatos para o intercâmbio de dados são algumas delas, sendo que a revolução tecnológica decorrente do uso de computadores trouxe novos meios para o alcance dos objetivos da organização da informação e forçou uma reconsideração sobre a forma e a função dos registros bibliográficos (SVENONIUS, 2000, p. 107) e, conseqüentemente, dos registros de autoridade. Em 1991, Tillett escreveu que

A tecnologia disponível para criar e manter um catálogo tem influenciado muito os tipos de dispositivos de ligação incluídos no catálogo e prescritos nas regras de catalogação [...] O ambiente computadorizado deve nos oferecer ainda mais possibilidades e devemos selecionar cuidadosamente aquelas que proveem os vínculos [*links*] mais efetivos, os melhores caminhos para a informação nos futuros sistemas de informação. (TILLETT, 1991, p. 402, tradução nossa).

O surgimento da Web no início da década de 1990, sem dúvida, significou uma revolução no modo de compartilhar informações. Os catálogos digitais, com seus dados bibliográficos e de autoridade, passaram a ser disponibilizados na Web e, apesar do novo ambiente tecnológico, os instrumentos de descrição, regras de catalogação e formatos para o intercâmbio de dados, principalmente, mantiveram-se em certa medida imutáveis, contribuindo, assim, para uma reprodução na Web do mesmo modelo de dados utilizado nos catálogos digitais anteriores a ela.

Partindo dos conceitos de Web Semântica e de *Linked Data*, introduzidos na primeira década do século XXI, a Web tem hoje entre suas características a busca por tecnologias para o relacionamento e o reuso de dados através de distintos espaços. Neste contexto, surge uma demanda pela disponibilização de dados bibliográficos e de autoridade utilizando as tecnologias da Web Semântica e as práticas de *Linked Data*, de modo que os catálogos possam beneficiar-se do relacionamento e do reuso de dados e possam beneficiar, com seus dados, outros espaços e serviços de informação.

Diante dessa demanda, o objetivo geral desta pesquisa, apresentado na Seção 1.2, é propor, a partir do conceito das funções dos dados de autoridade nos catálogos e seus benefícios nos ambientes de *Linked Data*, um modelo de publicação de dados de autoridade como *Linked Data* compreendendo políticas, procedimentos e tecnologias.

Para a proposição desse modelo, no entanto, são necessários: (1) o entendimento das funções dos dados de autoridade nos catálogos bibliográficos, entendimento esse alcançado por meio dos resultados apresentados neste capítulo; (2) o entendimento das tecnologias da Web Semântica, que é o objetivo do capítulo seguinte; e (3) a análise das iniciativas para a publicação de dados de autoridade como *Linked Data*, que é apresentado no Capítulo 4.

3 Web Semântica, suas tecnologias e *Linked Data*

Do desenvolvimento e da implementação de algumas das tecnologias da Web Semântica surgiu o conceito de *Linked Data*. Uma vez que o entendimento desse conceito e das principais tecnologias da Web Semântica a ele relacionadas é uma das bases para o desenvolvimento do modelo para a publicação de dados de autoridade como *Linked Data*, este capítulo destina-se a abordar tal conceito e tais tecnologias.

O surgimento da Web no início da década de 1990 trouxe novos meios para a publicação, o acesso e o intercâmbio de dados em ambientes digitais. A partir da proposta de criação de uma Web Semântica, apresentada em 2001 por Berners-Lee, Hendler e Lassila, tecnologias têm sido desenvolvidas e/ou aprimoradas para dar suporte ao acesso e ao processamento de dados presentes na Web, não somente pela sintaxe desses dados, mas também por sua semântica.

Ao apresentarem a proposta da Web Semântica, Berners-Lee, Hendler e Lassila (2001, p. 25, tradução nossa) destacam que a maior parte do conteúdo da Web é projetada para a leitura por humanos e não para que programas de computador possam manipulá-la a partir de seu significado e que a “Web Semântica trará estrutura para o conteúdo significativo das páginas da Web, criando um ambiente no qual agentes de softwares, percorrendo de uma página a outra, poderão rapidamente realizar tarefas sofisticadas para os usuários.” (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 25, tradução nossa).

Para este fim, os autores salientam que os “[...] computadores devem ter acesso a coleções de informações estruturadas e a conjuntos de regras que eles possam utilizar para realizar raciocínio automático” (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 26, tradução nossa). A estruturação dessas coleções de informações ocorre não somente pela sintaxe dos dados, mas também por sua semântica:

Para isso, não é suficiente armazenar os dados em uma sintaxe processável por máquina – nesse sentido, toda página HTML na Web é processável por máquina – é necessário que esses dados sejam dotados de uma semântica formal que claramente especifique quais conclusões devem ser obtidas a partir da informação coletada. (HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 11, tradução nossa).

Sobre isso, Heath e Bizer (2011, item 1.2.1, tradução nossa) entendem que “um fator chave na reusabilidade dos dados é a extensão na qual eles estão estruturados.

Quanto mais regular e bem definida a estrutura dos dados, mais facilmente as pessoas poderão criar ferramentas para processá-los de forma confiável para o seu reuso.” Os autores acrescentam que

Assim como a World Wide Web revolucionou o modo como conectamos e consumimos documentos, ela pode revolucionar o modo como descobrimos, acessamos, integramos e utilizamos os dados. A Web é o meio ideal para permitir esses processos devido a sua ubiquidade, a sua natureza distribuída e escalável e ao seu maduro e bem compreendido conjunto de tecnologias. (HEATH; BIZER, 2011, item 1.1, tradução nossa).

Cabe destacar que a Web Semântica diz respeito não somente aos dados sobre os recursos disponíveis na Web, por exemplo, um texto publicado em um site de notícias, uma foto em uma rede social ou um vídeo em um site de compartilhamento de vídeos, mas também sobre aqueles recursos que não estão disponíveis na Web, por exemplo, o autor do texto, o local retratado na fotografia ou os conceitos tratados no vídeo.

Para Hitzler, Krötzsch e Rudolph (2010, p. 1-2, tradução nossa) a Web Semântica é construída sobre três bases conceituais que envolvem os modelos para representação da realidade (ou do conhecimento), os esforços para a criação de uma inteligência artificial e a infraestrutura tecnológica já existente na Web:

- Construção de modelos: a busca pela descrição do mundo em termos abstratos para permitir um fácil entendimento de uma realidade complexa.
- Computação com conhecimento: o esforço para a construção de máquinas de raciocínio que possam obter conclusões significativas a partir de conhecimento codificado.
- Troca de informações: a transmissão de complexos recursos informacionais entre computadores que nos permite distribuir, interligar e reconciliar conhecimento em uma escala global.

Coyle (2016, p. 34, tradução nossa) destaca que “a Web Semântica enfatiza o crescimento de conhecimento e as interconexões entre os dados que podem vir de diferentes ambientes” e que, “embora esteja sendo utilizada em aplicações comerciais, a Web Semântica está mais orientada à descoberta e ao aumento de conhecimento do que ao controle [sobre ele]”. Sobre a relação entre Web Semântica e conhecimento, é necessário destacar que, pela natureza de suas tecnologias, a Web Semântica lida com conhecimento inferencial.

Para Santos e Alves (2009),

[...] a implementação da Web Semântica requer o trabalho conjunto das várias ferramentas tecnológicas e proporcionará em pequena, média e grande escala a tão necessária estruturação e representação informacional dos recursos e conseqüentemente sua melhor recuperação.

Berners-Lee, Hendler e Lassila (2001, p. 26, tradução nossa) sintetizam que o desafio da Web Semântica é “[...] prover uma linguagem que expresse os dados e as regras para o raciocínio sobre os dados e que permita que as regras de qualquer sistema existente de representação do conhecimento possam ser exportadas para a Web.”

Com o intuito de apoiar a construção de uma Web Semântica, o *World Wide Web Consortium* (W3C), consórcio com a missão de desenvolver de forma conjunta padrões para a Web, tem criado uma série de tecnologias voltadas à Web Semântica. Sobre essas tecnologias, Santarem Segundo (2015, p. 223-224) aponta que

Desde a proposta inicial de 2001, a Web Semântica vem ganhando força e agregando novas tecnologias, funcionalidades e evoluindo para tornar real o processo de construção de ambientes semânticos. Tecnologias como RDF (Resource Description Framework), XML (eXtensible Markup Language), OWL (Web Ontology Language) e todos os conceitos que as envolvem, ganham novas versões, descritos com clareza no W3C (World Wide Web Consortium), e tornam possível a materialização do conceito da Web Semântica. As tecnologias citadas estão diretamente relacionadas ao processo de construção da informação e armazenamento das mesmas, constituindo assim ambientes que possam ter conjunto de dados ligados semanticamente.

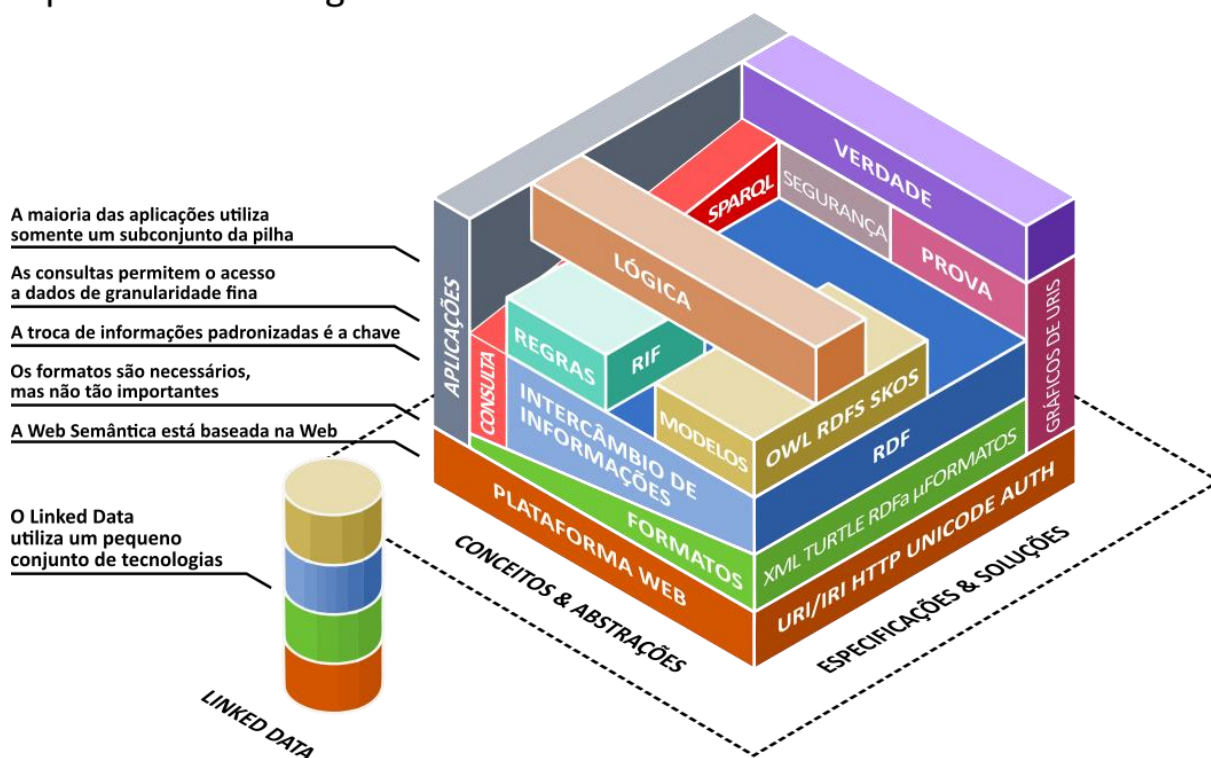
Hitzler, Krötzsch e Rudolph (2010, p. 1, tradução nossa) pontuam que

[...] similar à Web tradicional, a base das tecnologias da Web Semântica são os formatos de dados que podem ser utilizados para codificar conhecimento para processamento (de seus aspectos relevantes) em sistemas de computador, embora o foco esteja em diferentes formas de conhecimento.

As tecnologias da Web Semântica são frequentemente retratadas em camadas de acordo com suas funções. Um destes retratos é apresentado na Figura 9.

Figura 9 – Tecnologias da Web Semântica

A pilha de tecnologias da Web Semântica



Fonte: Nowack (2009, tradução nossa).

Em 2006, Berners-Lee, partindo da compreensão de que “a Web Semântica não é apenas sobre colocar dados na web. É sobre estabelecer relacionamentos [*links*] para que uma pessoa ou máquina possa explorar a web dos dados” (BERNERS-LEE, 2006, tradução nossa), introduziu o conceito de *Linked Data* ao propor quatro regras, ou melhores práticas²⁵, para que os dados publicados na Web sejam considerados dados vinculados, no inglês, *Linked Data*:²⁶

1. Use URIs como nomes [identificadores] para as coisas;
2. Use URIs HTTP, assim as pessoas poderão acessar esses nomes;

²⁵ Berners-Lee (2006, tradução nossa) destaca em seu texto: “Me referirei aos passos acima como sendo regras, mas eles são expectativas de comportamento. Infringí-las não arruinará nada, mas causará a perda de uma oportunidade de tornar os dados interconectados”. Nesse sentido, prefere-se o termo “melhores práticas” (HEATH; BIZER, 2011) ao invés do termo “regras”.

²⁶ O termo *Linked Data* é grafado de duas principais formas: *Linked Data* e *linked data*. Uma vez que não foi localizada na literatura uma distinção entre essas duas formas, nesta pesquisa é utilizada a forma *Linked Data*. Nas publicações em língua portuguesa, *Linked Data* é traduzido de diferentes formas, entre elas dados “lincados”, dados vinculados, dados interligados, dados ligados e dados ligados entre si. Na falta de uma tradução consensual e por questões de visibilidade internacional, nesta pesquisa é mantido o termo original em língua inglesa.

3. Quando alguém acessar um URI, forneça informação útil utilizando os padrões (RDF, SPARQL);
4. Inclua *links* para outros URIs. Assim poderão descobrir mais coisas.

Segundo van Hooland e Verborgh (2015, p. 127, destaque dos autores, tradução nossa),

Após alguns anos, a web semântica tornou-se sinônimo de complexidade e sua adoção foi lenta. Isso nitidamente contrastou com a facilidade com a qual a web foi adotada apenas uma década antes. A web semântica foi vista mais ou menos como um parque de diversão dos pesquisadores da inteligência artificial, enquanto as pessoas em geral, na verdade, desejavam apenas algo que “funcionasse”. Publicar dados e criar aplicações tornou-se muito complexo para não especialistas, assim somente poucos conjuntos de dados foram disponibilizados.

o que, segundo eles, foi inspiração para os princípios de *Linked Data* propostos por Berners-Lee:

Em vez de preocupar-se muito com os conceitos subjacentes, o linked data prioriza a obtenção de dados em formato legível por máquina, enfatizando a importância de links para outros conceitos para criar significado. (VAN HOOLAND; VERBORGH, 2015, p. 127, tradução nossa).

Por utilizar as tecnologias desenvolvidas para a Web (e para a Web Semântica) as melhores práticas para a publicação de *Linked Data* são construídas “[...] diretamente sobre a arquitetura da Web e aplicam essa arquitetura na tarefa de compartilhar dados em uma escala global” (HEATH; BIZER, 2011, item 2, tradução nossa). Para Heath e Bizer (2011, item 1.3, tradução nossa),

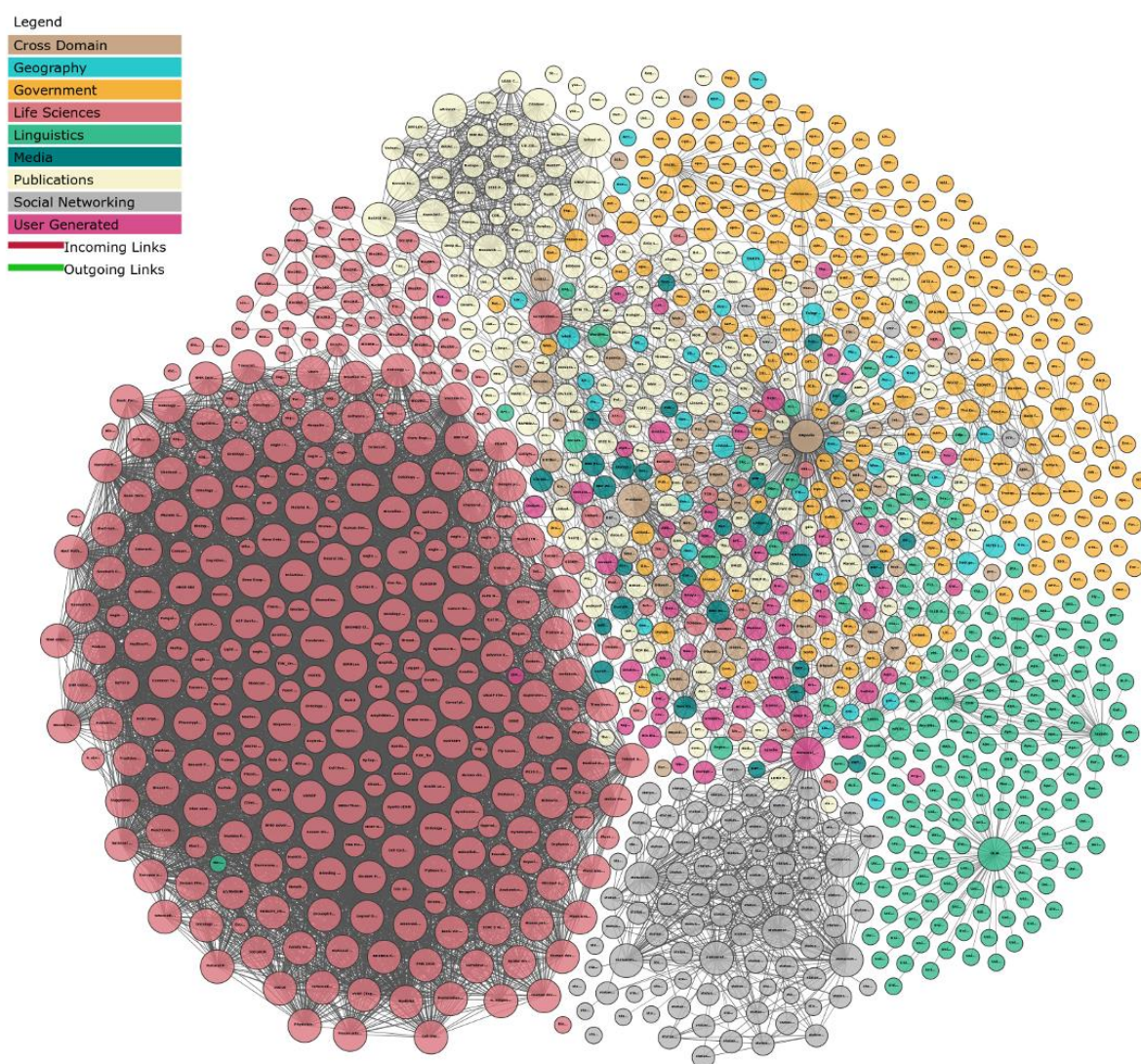
Assim como os *hiperlinks* na Web clássica conectam documentos em um espaço de informação global único, *Linked Data* permite que *links* sejam estabelecidos entre diferentes itens de diferentes fontes de dados e, assim, permite conectar essas fontes em um único espaço de dados global. O uso de padrões da Web e de um modelo de dados comum torna possível a implementação de aplicações genéricas que operem sobre todo o espaço de dados. Esta é a essência do *Linked Data*.

Sobre esses *links*, Marcondes (2012, p. 173, grifo do autor) explica que são *links* semânticos, com “significado” para os programas:

Ao contrário, *links* convencionais nada mais são (além de uma eventual etiqueta textual significativa para usuários humanos) que meios para que programas navegadores, a partir de um recurso, acessem outro, *sem explicitar qual o significado da ligação entre os recursos*. Sendo significativos para programas, links semânticos podem ser processados de forma mais rica por eles, explorando e enriquecendo cognitivamente o significado (legível por máquina) da ligação entre ambos os recursos.

Para a construção deste “espaço de dados global”, considera-se nesta pesquisa que dois principais tipos de dados têm sido publicados utilizando tecnologias da Web Semântica e as melhores práticas indicadas por Berners-Lee: (1) conjuntos de dados descrevendo recursos, os chamados *datasets*, e (2) vocabulários descrevendo as propriedades e as classes utilizadas na descrição de recursos. Em outras palavras, “coleções de informações estruturadas” e “regras [...] para realizar raciocínio automático” (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 26, tradução nossa).

Figura 10 – *Linking Open Data cloud diagram*



Fonte: Abele (2017).

Conjuntos de dados, assim como vocabulários, têm sido publicados em diversos domínios. A publicação de conjuntos de dados tem sido acompanhada por Abele et al. (2017), que reúnem informações sobre esses conjuntos e sobre as relações entre eles e as apresentam na forma de um gráfico com nove categorias (dados sobre publicações, dados sobre ciências da vida, dados gerais entre domínios, dados de redes sociais, dados geográficos, dados governamentais, dados sobre mídias, dados sobre conteúdos criados pelos usuários, e dados linguísticos), o *Linking Open Data cloud*, apresentado na Figura 10.

Desde sua introdução em 2006, o conceito de *Linked Data* tem evoluído e, hoje, já é encontrado na literatura um conceito dele derivado: *Linked Open Data* (LOD). Berners-Lee (2010, tradução nossa) define *Linked Open Data* como “[...] *Linked Data* que é disponibilizado sob uma licença aberta que não impeça seu livre reuso” e apresenta um sistema de classificação para os dados disponibilizados na Web em que quanto maior a quantidade de estrelas mais vinculados e abertos são os dados:

- ★ Disponível na Web (em qualquer formato), mas como uma licença aberta [*open license*], para ser *Open Data*;
- ★★ Disponível como dado estruturado e legível por máquina (por exemplo, uma planilha do Excel ao invés de uma imagem digitalizada de uma tabela);
- ★★★ Assim como no (2), mas em um formato não proprietário (por exemplo, no formato CSV ao invés do formato do Excel);
- ★★★★ Todos os anteriores utilizando padrões abertos do W3C (RDF e SPARQL) para identificar coisas, assim as pessoas poderão referir-se as suas coisas;
- ★★★★★ Todos os anteriores e: relacionar seus dados aos dados de outras pessoas para prover contexto (BERNERS-LEE, 2010, tradução nossa).

Cabe aqui destacar que o conceito de *Linked Data* refere-se à interoperabilidade técnica dos dados, enquanto que o de *Linked Open Data* tem seu foco na interoperabilidade legal desses dados, não sendo obrigatório que o todo *Linked Data* seja aberto (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Destaca-se também que *Linked Data* e Web Semântica não competem entre si, ao contrário, “[...] o *Linked Data* pode ser considerado as blocos que constituem a web semântica” (VAN HOOLAND; VERBORGH, 2015, p. 127, tradução nossa).

Ao apresentar as tecnologias da Web Semântica, a Figura 9 traça um recorte destacando quais dessas tecnologias relacionam-se diretamente ao conceito de *Linked Data*. Para o entendimento desse conceito, são descritas neste capítulo as tecnologias a ele relacionadas: na Seção 3.1 são abordados os URIs HTTP, que se relacionam às melhores práticas 1 e 2; na Seção 3.2 é abordado o RDF, que se relaciona à prática 3; e na Seção 3.3 são abordados o RDFS e a OWL, que se relacionam à prática 4. No decorrer dessas seções é discutido como essas tecnologias relacionam-se a cada uma das melhores práticas para a publicação de *Linked Data*.

3.1 Uniform Resource Identifier (URI) Hypertext Transfer Protocol (HTTP)

A representação de um recurso demanda primeiramente sua identificação (FERREIRA; SANTOS, 2013, p. 17), sendo que, para o propósito de construção de uma Web Semântica, essa identificação deve dar-se de modo compreensível às máquinas (COYLE, 2010, p. 21).

Essa identificação ocorre por meio de identificadores e, embora haja uma variedade de tipos de identificadores, para a Web Semântica é necessário o uso de identificadores do tipo *Uniform Resource Identifier (URI)*²⁷ (COYLE, 2010, p. 21). No contexto de *Linked Data*, Berners-Lee (2006) aponta como primeira regra a necessidade do uso de URIs para a identificação dos recursos, estejam eles disponíveis ou não na Web, os chamados objetos do mundo real ("*real-world objects*") ou coisas ("*things*") (WORLD WIDE WEB CONSORTIUM, 2008).

URIs englobam os *Uniform Resource Locators* - URL, isto é, os endereços Web que são utilizados para acessar os recursos online. Cada URL é um URI válido e, portanto, pode ser utilizado como um identificador em documentos que descrevem recursos disponíveis na Web. (FERREIRA; SANTOS, 2013, p. 18)

Ferreira e Santos (2013, p. 18) esclarecem que quando um URI é utilizado para identificar um recurso não disponível na Web, por exemplo, uma pessoa, uma cidade ou um livro impresso, seu propósito é apenas o de identificar esse recurso, diferentemente

²⁷ Em 2005, o padrão URI foi complementado pelo *Internationalized Resource Identifier (IRI)*. Os IRIs permitem o uso de um conjunto de caracteres maior que o utilizado nos URIs. No entanto, nesta pesquisa manteve-se o uso de URI, uma vez que esse termo ainda é encontrado com maior frequência na literatura consultada.

de um URI referente a um recurso disponível na Web, que tem o propósito de localizar um recurso ou prover acesso a ele, além de identificá-lo.

Os URIs podem ser de diferentes esquemas, tais como HTTP, FTP, DOI, ISBN e ISSN. Como segunda melhor prática para a publicação de *Linked Data*, Berners-Lee (2006) descreve o uso de URIs do tipo (esquema) HTTP, que apresentam vantagens em relação aos URIs de outros tipos.

Primeiramente, os URIs HTTP são identificadores únicos e globais que são criados de forma simples e descentralizada, uma vez que qualquer proprietário de um domínio Web pode criar novos URIs (HEATH; BIZER, 2011).

Em segundo lugar, além de identificar os recursos, os URIs HTTP são utilizados também para recuperar informações sobre eles (HEATH; BIZER, 2011), o que pode ocorrer a partir da negociação de conteúdo, uma funcionalidade existente na Web. Quando o propósito de um URI é apenas o de identificar um recurso – e não o de acessá-lo – seu uso em um navegador não necessariamente recuperará alguma informação sobre o recurso. Por exemplo, o URI *http://fabricioassumpcao.com/uri* pode ser utilizado para identificar a pessoa “Fabrício Silva Assumpção”; ao inserirmos esse URI em um navegador nenhuma informação sobre essa pessoa será recuperada. No entanto, com a negociação de conteúdo, as informações sobre o recurso identificado pelo URI podem ser recuperadas, inclusive em diferentes formatos, se necessário.

Na negociação de conteúdo, o servidor e a aplicação que está acessando o URI trocam dados sobre os formatos nos quais a descrição do recurso está disponível, sobre possíveis redirecionamentos e sobre os formatos preferidos pela aplicação para receber tal descrição. Se o URI está sendo acessado por uma aplicação como um navegador, por exemplo, o servidor poderá retornar uma página HTML com a descrição do recurso. Se o URI está sendo acessado por uma aplicação que utiliza tecnologias da Web Semântica, por exemplo, um navegador semântico²⁸, o servidor poderá retornar a descrição do recurso em XML, JSON-LD, *Turtle*, etc. Essa negociação de conteúdo através de um serviço de resolução (*dereferencing service*) faz com que um mesmo URI possa ser utilizado para recuperar tanto informações destinadas aos humanos, por exemplo, uma página HTML, quanto dados estruturados para o processamento por máquinas, por

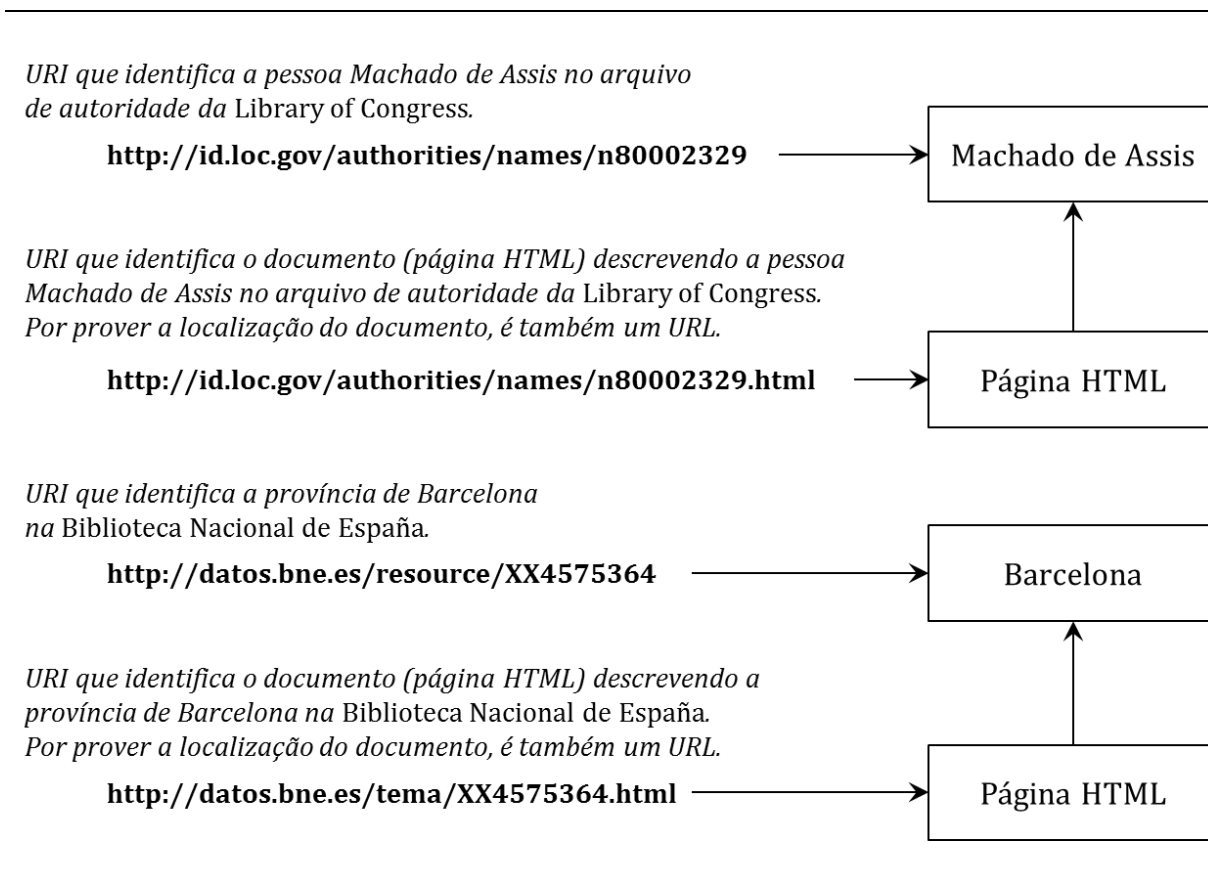
²⁸ O W3C mantém duas listas contendo navegadores semânticos, disponíveis em: https://www.w3.org/2001/sw/wiki/Category:Semantic_Web_Browser e https://www.w3.org/2001/sw/wiki/Category:RDF_or_OWL_Browser. Acesso em: 12 dez. 2017.

exemplo, um documento XML (HEATH; BIZER, 2011; HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 36; WILLER; DUNSIRE, 2013, p. 143, 180-181).

Um exemplo de negociação de conteúdo é observado a partir do URI <http://id.loc.gov/authorities/names/n80002329> que identifica a pessoa “Machado de Assis” no arquivo de autoridade da *Library Congress*. Se utilizado em um navegador, este URI retornará a página HTML <http://id.loc.gov/authorities/names/n80002329.html>, que contém a descrição da pessoa. Se utilizado em um navegador semântico, por exemplo, o mesmo URI poderá retornar documentos XML (<http://id.loc.gov/authorities/names/n80002329.rdf>), N-Triples (<http://id.loc.gov/authorities/names/n80002329.nt>) e JSON-LD (<http://id.loc.gov/authorities/names/n80002329.json>), que contêm a descrição da pessoa em formatos processáveis por programas de computador.

Neste ponto, em se tratando de recursos não disponíveis na Web, é importante ressaltar a necessidade de distinção entre o URI que identifica o recurso em si e o URI que identifica o documento que descreve o recurso na Web (HEATH; BIZER, 2011). Uma das formas de apresentar essa distinção é denominando os recursos não disponíveis na Web como entidades ou objetos do mundo real, e aqueles disponíveis como documentos da Web (WORLD WIDE WEB CONSORTIUM, 2008). Um exemplo essa distinção é apresentado na Figura 11.

Figura 11 – URIs identificando objetos do mundo real e documentos



Fonte: Elaborada pelo autor.

Essa distinção, que se dá por meio da atribuição de diferentes URIs, permite que as declarações sobre o recurso não sejam ambíguas, por exemplo, a pessoa nasceu (ou seja, foi criada) em 21 de junho de 1839, enquanto que o documento que a descreve foi criado em 12 de fevereiro de 2016. As declarações sobre os recursos serão abordadas na Seção 3.2.

Sobre o uso de URIs, Coyle (2016, p. 55, destaque do autor, tradução nossa) ressalta que eles são destinados às máquinas e não aos humanos:

Ninguém deseja ler, e muito menos digitar, “<http://id.loc.gov/authorities/subjects/sh85038796>” para o cabeçalho de assunto “*Dogs*” na *Library of Congress*. Todos os identificadores podem ter rótulos legíveis por humanos e o pressuposto é que em todas as situações em que um humano estiver interagindo com os dados esse rótulo seja exibido. [...] Assim, um catalogador escolherá um cabeçalho de assunto, por exemplo “*Dogs in literature*”, de uma lista e o dado armazenado será “<http://id.loc.gov/authorities/subjects/sh85038823>”.

Em síntese, os URIs, que são utilizados para identificar recursos disponíveis ou não na Web e recuperar informações sobre eles, são essenciais para a Web Semântica e

para a publicação de *Linked Data*. O uso de URIs, no entanto, por si só não garante a inclusão de semântica e nem de *links* entre os dados, o que é obtido a partir da utilização de tecnologias como o *Resource Description Framework* (RDF), apresentado na seção seguinte.

3.2 *Resource Description Framework* (RDF)

A terceira prática para a publicação de *Linked Data* listada por Berners-Lee (2006) é “quando alguém acessar um URI, forneça informação útil utilizando os padrões (RDF, SPARQL)”. Os URIs e o modo de prover informações sobre o que eles identificam – a negociação de conteúdo – foram abordados na seção anterior. O *Resource Description Framework* (RDF), um modelo de dados para a organização dessas informações, é descrito nesta seção.

O W3C define o RDF como “[...] uma estrutura para a expressão de informações sobre recursos” em que “recursos podem ser quaisquer coisas, incluindo documentos, pessoas, objetos físicos e conceitos abstratos” (WORLD WIDE WEB CONSORTIUM, 2014a, tradução nossa). Para Hitzler, Krötzsch e Rudolph (2010, p. 19, tradução nossa), o RDF

[...] é uma linguagem formal para a descrição de informações estruturadas. O objetivo do RDF é permitir que aplicações troquem dados na Web preservando seu significado original. Em oposição à HTML e à XML, a principal intenção agora não é exibir os documentos corretamente, mas sim permitir o posterior processamento e recombinação das informações contidas neles. O RDF, conseqüentemente, é frequentemente visto como o formato básico de representação para o desenvolvimento da Web Semântica.

O RDF é apresentado em um conjunto de documentos publicados pelo W3C: *RDF Primer*,²⁹ que apresenta uma introdução ao RDF; *RDF Concepts and Abstract Syntax*,³⁰ que descreve o modelo de dados RDF e os conceitos relacionados a ele; *RDF Semantics*,³¹ que contém a especificação técnica do RDF e do *RDF Schema*; e os documentos que especificam a serialização do RDF em formatos específicos, que são apresentados mais

²⁹ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 Primer**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>>. Acesso em: 12 dez. 2017.

³⁰ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 Concepts and Abstract Syntax**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>>. Acesso em: 12 dez. 2017.

³¹ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 Semantics**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/REC-rdf11-mt-20140225/>>. Acesso em: 12 dez. 2017.

adiante nesta seção. Considerando que o RDF não é o objeto central deste estudo, uma breve apresentação dele é conduzida nesta seção com ênfase no modelo de dados e nos principais formatos de serialização, uma vez que estes aspectos relacionam-se diretamente ao modelo para a publicação de dados de autoridade como *Linked Data* desenvolvido nesta pesquisa.³²

No modelo de dados RDF, as informações sobre os recursos são descritas por meio de declarações, denominadas triplas, compostas por três partes: recurso, propriedade e valor.³³ As declarações são comparadas às orações em linguagem natural, assim, a oração “Machado de Assis nasceu no Rio de Janeiro” teria como recurso “Machado de Assis”, como propriedade “nasceu no” e como valor “Rio de Janeiro”.

Uma vez que cada declaração inclui uma propriedade do recurso, múltiplas declarações podem ser feitas sobre um mesmo recurso, por exemplo, “Machado de Assis tem o nome completo Joaquim Maria Machado de Assis”, “Machado de Assis nasceu em 21 de junho de 1839”, “Machado de Assis foi casado com Carolina Augusta Xavier de Novais”, etc.

O valor de uma declaração é outro recurso ou uma literal. O valor é considerado um recurso quando é ou pode ser utilizado como recurso em outra declaração. Na declaração “Machado de Assis foi casado com Carolina Augusta Xavier de Novais”, por exemplo, o valor é outro recurso, uma vez que ele é ou pode ser o recurso de outras declarações: “Carolina Augusta Xavier de Novais nasceu no Porto”, “Carolina Augusta Xavier de Novais morreu em 1904”, etc.

Nos casos em que o valor da declaração não é um recurso, diz-se que é um valor literal ou valor de dado. Valores literais incluem, por exemplo, conjuntos de caracteres alfanuméricos na forma de palavras, frases, datas, valores numéricos, valores lógicos booleanos, etc. Nas declarações “Machado de Assis tem o nome completo Joaquim Maria Machado de Assis” e “Carolina Augusta Xavier de Novais morreu em 1904” os valores “Joaquim Maria Machado de Assis” e “1904” são valores literais.

Uma vez que o RDF destina-se ao processamento por sistemas computacionais, é necessário que os componentes das declarações sejam identificados de modo apropriado a esses sistemas. Como abordado na seção anterior, os URIs do tipo HTTP

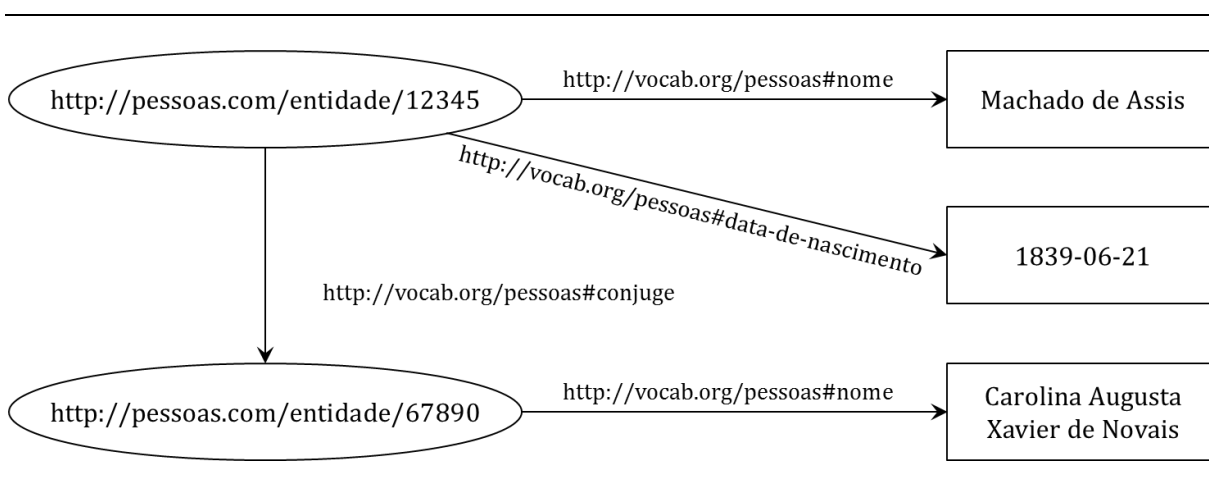
³² Descrições mais detalhadas do RDF são encontradas em Catarino e Souza (2012) e em Ferreira e Santos (2013).

³³ Na literatura são encontrados também os termos sujeito, predicado e objeto que correspondem, respectivamente, a recurso, propriedade e valor.

são os indicadores para a identificação de recursos no contexto da Web Semântica e do *Linked Data*. Assim, os componentes das declarações RDF (recursos e propriedades) devem ser identificados por URIs, sendo uma exceção os valores literais que, por se tratarem de cadeias de caracteres (*strings*), não são identificados por URIs (WILLER; DUNSIRE, 2013, p. 101).

No modelo RDF, as declarações são representadas por grafos em que: (1) os recursos são representados por elipses, (2) as propriedades são representadas por setas que partem dos recursos e apontam para os valores, e (3) os valores são também representados por elipses (quando são recursos de outras declarações) ou por retângulos (quando são valores literais). A Figura 12 apresenta um grafo RDF com algumas das declarações utilizadas anteriormente como exemplo.³⁴

Figura 12 – Grafo RDF



Fonte: Elaborada pelo autor.

Cabe enfatizar que, no RDF, tudo que é identificado por um URI é um recurso, mesmo que, em uma determinada tripla não ocupe a posição de recurso. A propriedade de uma tripla, por exemplo, é o recurso de outra tripla. As declarações feitas sobre uma propriedade, que têm por objetivo a criação de vocabulários de propriedades, são abordadas na Seção 3.3. Uma ou mais declarações em RDF formam um conjunto de dados, também referido como um grafo RDF.³⁵

³⁴ Na Figura 12, assim como nos demais exemplos apresentados nesta seção, os recursos e as propriedades são identificados por URIs fictícios.

³⁵ Sobre a diferença entre os termos “conjunto de dados” e “grafo”, Alexander et al. (2009, p. 2-3, tradução nossa) entendem “[...] um conjunto de dados como uma coleção significativa de triplas que lidam com

Além da forma de grafos, as declarações em RDF podem ser apresentadas seguindo um formato de serialização. A serialização, entendida aqui como a “transformação de estruturas de dados complexas para cadeias de caracteres lineares” (FERREIRA; SANTOS, 2013, p. 19), destina-se ao processamento das declarações por sistemas computacionais, enquanto que a apresentação por meio de grafos volta-se à compreensão do RDF por parte de seres humanos (FERREIRA; SANTOS, 2013, p. 19; WILLER; DUNSIRE, 2013, p. 173). Sobre o uso de grafos, Ferreira e Santos (2013, p. 19) destacam que

Para humanos, entender grafos visualmente funciona muito bem se os grafos são pequenos, com poucos nós, arestas e literais. Mas os gigantescos bancos de dados atuais contêm milhares de dados que gerariam grafos absurdamente grandes.

Alguns dos formatos para a serialização do RDF são exemplificados na Figura 13 que apresenta, nesses formatos, as mesmas declarações do grafo da Figura 12.

Figura 13 – Formatos para a serialização do RDF

Extensible Markup Language (XML)
<pre><?xml version="1.0" encoding="utf-8" ?> <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:vocabp="http://vocab.org/pessoas#"> <rdf:Description rdf:about="http://pessoas.com/entidade/12345"> <vocabp:nome>Machado de Assis</vocabp:nome> <vocabp:data-de-nascimento>1839-06-21</vocabp:data-de-nascimento> <vocabp:conjuge> <rdf:Description rdf:about="http://pessoas.com/entidade/67890"> <vocabp:nome>Carolina Augusta Xavier de Novais</vocabp:nome> </rdf:Description> </vocabp:conjuge> </rdf:Description> </rdf:RDF></pre>
Notation 3 (N3)
<pre>@prefix vocabp: <http://vocab.org/pessoas#> . <http://pessoas.com/entidade/12345> vocabp:nome "Machado de Assis" ; vocabp:data-de-nascimento "1839-06-21" ;</pre>

certo tópico, têm origem em certa fonte ou processo, estão hospedadas em certo servidor ou são agregadas por certo custodiador. O termo, assim, possui uma dimensão social que não é fácil de ser capturada em uma definição formal. Isso diferencia conjuntos de dados de grafos RDF, que são constructos puramente matemáticos. Qualquer conjunto arbitrário de triplas em RDF é, por definição, um grafo RDF, independentemente da semântica das triplas”.

```
vocabp:conjuge <http://pessoas.com/entidade/67890> .
<http://pessoas.com/entidade/67890> vocabp:nome "Carolina Augusta Xavier de Novais" .
```

N-Triples

```
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#nome> "Machado de Assis" .
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#data-de-nascimento> "1839-06-21" .
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#conjuge>
<http://pessoas.com/entidade/67890> .
<http://pessoas.com/entidade/67890> <http://vocab.org/pessoas#nome> "Carolina Augusta Xavier de
Novais" .
```

Terse RDF Triple Language (Turtle)

```
@prefix vocabp: <http://vocab.org/pessoas#> .

<http://pessoas.com/entidade/12345>
  vocabp:nome "Machado de Assis";
  vocabp:data-de-nascimento "1839-06-21";
  vocabp:conjuge <http://pessoas.com/entidade/67890> .
<http://pessoas.com/entidade/67890> vocabp:nome "Carolina Augusta Xavier de Novais" .
```

JavaScript Object Notation for Linked Data (JSON-LD)

```
{
  "@context": {
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "vocabp": "http://vocab.org/pessoas#",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@graph": [
    {
      "@id": "http://pessoas.com/67890",
      "vocabp:nome": "Carolina Augusta Xavier de Novais"
    },
    {
      "@id": "http://pessoas.com/12345",
      "vocabp:conjuge": {
        "@id": "http://pessoas.com/67890"
      },
      "vocabp:data-de-nascimento": "1839-06-21",
      "vocabp:nome": "Machado de Assis"
    }
  ]
}
```

Fonte: Elaborada pelo autor.

Originalmente, o W3C propôs a serialização do RDF utilizando a *Extensible Markup Language* (XML).³⁶ Posteriormente, foi introduzido o formato *Notation 3* (N3),³⁷

³⁶ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 XML Syntax**. Cambridge, 2014. Disponível em: <https://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/>. Acesso em: 12 dez. 2017.

que, em comparação à XML, é mais legível aos humanos (WILLER; DUNSIRE, 2013, p. 173-174).

A segunda especificação do RDF, publicada em 2004, trouxe como possibilidade para a serialização o formato *N-Triples*,³⁸ criado a partir do N3. Informalmente, o *N-Triples* foi expandido por seus usuários, recebendo o nome de *Turtle* (HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 25-26). Este formato permaneceu informal até 2011, quando foi reconhecido pelo W3C.³⁹ Mais recentemente, o W3C tem recomendado também a utilização do formato *JavaScript Object Notation for Linked Data* (JSON-LD)⁴⁰ considerando a aceitação do JSON em diversas linguagens de programação, como destacam Heath e Bizer (2011, tradução nossa):

A disponibilidade de uma serialização do RDF em JSON é altamente desejável, dado o crescente número de linguagens de programação com suporte nativo ao JSON, incluindo aquelas principais para a programação Web, tais como *JavaScript* e PHP. Assim, a publicação de dados RDF em JSON torna os dados acessíveis aos desenvolvedores Web sem a necessidade de instalar bibliotecas de softwares adicionais para a análise e a manipulação dos dados RDF.

Considerando a facilidade de compreensão da serialização utilizando *Turtle*, tal formato é utilizado nesta pesquisa para a apresentação de exemplos. As principais características da serialização do RDF com o *Turtle* estão ilustradas na Figura 14.

Figura 14 – Serialização do RDF utilizando o formato *Turtle*

Fragmento 1

```
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#nome> "Machado de Assis" .
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#data-de-nascimento> "1839-06-21" .
<http://pessoas.com/entidade/12345> <http://vocab.org/pessoas#conjuge>
<http://pessoas.com/entidade/67890> .
<http://pessoas.com/entidade/67890> <http://vocab.org/pessoas#nome> "Carolina Augusta Xavier de
Novais" .
```

³⁷ WORLD WIDE WEB CONSORTIUM. **Notation3 (N3)**: a readable RDF syntax. Cambridge, 2011. Disponível em: <<https://www.w3.org/TeamSubmission/2011/SUBM-n3-20110328/>>. Acesso em: 12 dez. 2017.

³⁸ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 N-Triples**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/REC-n-triples-20140225/>>. Acesso em: 12 dez. 2017.

³⁹ WORLD WIDE WEB CONSORTIUM. **RDF 1.1 Turtle**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/REC-turtle-20140225/>>. Acesso em: 12 dez. 2017.

⁴⁰ WORLD WIDE WEB CONSORTIUM. **JSON-LD 1.0**. Cambridge, 2014. Disponível em: <<https://www.w3.org/TR/2014/REC-json-ld-20140116/>>. Acesso em: 12 dez. 2017.

Fragmento 2

```
<http://pessoas.com/entidade/12345>
  <http://vocab.org/pessoas#nome> "Machado de Assis" ;
  <http://vocab.org/pessoas#data-de-nascimento> "1839-06-21" ;
  <http://vocab.org/pessoas#conjuge> <http://pessoas.com/entidade/67890> .
```

Fragmento 3

```
<http://tesouro.com.br/conceitos/0026>
  <http://propriedades.org/tesouro#termo-autorizado> "Educação"@pt ;
  <http://propriedades.org/tesouro#termo-autorizado> "Education"@en ;
  <http://propriedades.org/tesouro#termo-autorizado> "Educación"@es .
```

Fragmento 4

```
<http://tesouro.com.br/conceitos/0026> <http://propriedades.org/tesouro#termo-autorizado>
"Educação"@pt , "Education"@en , "Educación"@es .
```

Fragmento 5

```
@prefix vocabg: <http://vocab.org/grupos#> .
@prefix vocabp: <http://vocab.org/pessoas#> .
@prefix local: <http://propriedades.org/locais#> .

<http://arquivodeautoridade.com/entidade/98765>
  vocabg:nome "Organização das Nações Unidas para a Educação, a Ciência e a Cultura"@pt ;
  vocabg:nome "United Nations Educational, Scientific and Cultural Organization"@en ;
  vocabg:sigla "UNESCO" ;
  vocabg:fundacao "1946" ;
  vocabg:diretor <http://pessoas.com/entidade/4321> ;
  vocabg:sede <http://locais.gov/fr0123456> .

<http://pessoas.com/entidade/4321> vocabp:nome "Irina Bokova" .

<http://locais.gov/fr0123456>
  local:nome "Paris" ;
  local:pais <http://locais.gov/fr> .

<http://locais.gov/fr>
  local:nome "França"@pt ;
  local:nome "France"@fr .
```

Fonte: Elaborada pelo autor.

Pela Figura 14, nota-se que na serialização do RDF utilizando *Turtle*:

- as partes de uma declaração (recurso, propriedade e valor) são separadas por um espaço (fragmento 1);
- os URIs são indicados entre os sinais de menor e de maior (fragmento 1);
- os valores literais são indicados entre aspas (fragmento 1);
- um ponto final indica o término de uma declaração (fragmento 1);

- as declarações sobre um mesmo recurso podem ser agrupadas por meio de um ponto e vírgula, sem a necessidade de repetição do URI do recurso (fragmento 2);
- o idioma de um valor literal pode ser indicado com o uso de um “@” seguido do código de identificação do idioma⁴¹ (fragmento 3);
- os diversos valores de uma mesma propriedade podem ser agrupados por meio de uma vírgula (fragmento 4);
- os *namespaces* podem ser abreviados utilizando prefixos; um *namespace* refere-se a um conjunto de URIs; os URIs *http://vocab.org/grupos#nome*, *http://vocab.org/grupos#sigla*, *http://vocab.org/grupos#fundacao*, *http://vocab.org/grupos#diretor*, por exemplo, fazem parte do *namespace* identificado pelo URI *http://vocab.org/grupos#*; este *namespace* pode ser abreviado por meio de um prefixo, por exemplo, *vocabg*, de modo que, ao realizar as declarações, todos os URIs deste *namespace* contarão com esse mesmo prefixo: *vocabg:nome*, *vocabg:sigla*, *vocabg:fundacao*, *vocabg:diretor*, etc. (fragmento 5).

Os demais aspectos técnicos deste formato estão contemplados em sua especificação e uma explicação sobre eles foge do escopo desta pesquisa.

Mesmo que as possibilidades de relacionar dados utilizando RDF assemelhem-se, num primeiro momento, às possibilidades oferecidas por uma base de dados construída sobre um modelo relacional, é necessário esclarecer que o contexto em que o RDF, as tecnologias da Web Semântica e as práticas de *Linked Data* estão inseridos possui diferenças em relação ao contexto de uma base de dados. Sobre essas diferenças, Coyle (2016, p. 54, tradução nossa) destaca que

Diferente da maioria dos modelos de dados que o precede, desde o modelo entidade-relacionamento até o orientado a objeto, o RDF não surge do mundo dos negócios do qual surgiram nossas melhorias tecnológicas anteriores. A Web Semântica, como o nome sugere, surge a partir da tecnologia da web. Esta é uma diferença significativa, por exemplo, em relação às tecnologias de bases de dados, pois a web é uma plataforma aberta e o espaço onde publicamos dados acessíveis, enquanto que as bases de dados são privadas e fechadas, mantidas dentro de empresas e frequentemente altamente controladas em termos de acesso. Isso significa que muitas das premissas que direcionam os padrões da Web Semântica são bem diferentes daquelas encontradas no processamento de dados de negócios.

⁴¹ Os códigos de identificação dos idiomas são estabelecidos a partir do documento *Tags for Identifying Languages*. Disponível em: <<https://tools.ietf.org/html/rfc5646>>. Acesso em: 12 dez. 2017.

A literatura (BAKER; COYLE; PETIYA, 2014, p. 565-566; COYLE, 2016; WILLER; DUNSIRE, 2013) destaca três premissas, intrinsecamente relacionadas, que direcionam o desenvolvimento dos padrões da Web Semântica e que são essenciais para uma melhor compreensão acerca do uso do RDF e das regras aplicadas aos dados na Web Semântica:

- **mundo aberto** (*open world*): um conjunto de declarações sobre um recurso está sempre incompleto, não importando a abrangência e o nível de detalhamento das declarações já feitas, assim, a ausência de uma declaração é compreendida como uma evidência de que a declaração ainda não foi feita, e não de que a declaração não se aplica ou de que ela é falsa (WILLER; DUNSIRE, 2013, p. 119);
- **nome não-único** (*Non-Unique Name Assumption* – NUNA): um recurso descrito em RDF pode ser identificado por mais de um URI; o autor Machado de Assis, por exemplo, é identificado por diferentes URIs: um no *Wikidata*, outro no arquivo de autoridade da LC e um terceiro no VIAF; assim, não se pode assumir que as coisas são diferentes só porque são identificadas por URIs diferentes (BAKER; COYLE; PETIYA, 2014, p. 566);
- **“qualquer um pode dizer qualquer coisa sobre qualquer coisa”** (“*anyone can say anything about anything*”): em consonância com a premissa de um mundo aberto, há sempre uma nova declaração que alguém pode fazer sobre um determinado recurso; a Web Semântica é projetada para operar em uma escala global e acomodar informações de qualquer fonte, suas tecnologias, portanto, são projetadas para acomodar múltiplas fontes de informação refletindo múltiplos pontos de vista, o RDF, por exemplo, permite a junção de múltiplas fontes ao utilizar URIs compartilhados entre essas fontes para alinhar as informações em um todo unificado (BAKER; COYLE; PETIYA, 2014, p. 565; WILLER; DUNSIRE, 2013, p. 119).

Embora os exemplos utilizados até aqui tenham apresentado conjuntos de declarações com diversas propriedades de um mesmo recurso (por exemplo, várias declarações sobre “Machado de Assis”), é necessário destacar que o RDF tem como unidade básica uma declaração, e não um conjunto de declarações. Ou seja, é possível criar uma única declaração sobre uma única propriedade de um recurso, sem a necessidade de declarações que descrevam as demais propriedades do recurso em questão.

Sobre essa característica do RDF – a possibilidade de trabalhar com declarações individuais – entende-se nesta pesquisa que ela:

- é uma das diferenças do RDF em relação a outros modelos, por exemplo, do modelo utilizado na construção de catálogos digitais, apresentado na Seção 2.3, que requer a existência de registros (bibliográficos, de autoridade, de item) descrevendo várias propriedades de um recurso de forma conjunta;
- decorre da necessidade de reunir declarações advindas de conjuntos de dados distintos, possibilitando o agrupamento de declarações sobre um mesmo recurso advindas de diversas fontes; e
- favorece a utilização de propriedades e classes de vocabulários distintos, ainda que dentro de um mesmo conjunto de dados, como é abordado na Seção 3.3.

Em síntese, o modelo de dados RDF provê um meio flexível de realizar declarações sobre recursos (descrevê-los e relacioná-los) (HEATH; BIZER, 2011). Operando de modo similar ao conjunto de associações realizadas na mente humana (SANTAREM SEGUNDO, 2015, p. 226), o RDF é entendido como um modelo aparentemente simples de coisas e de relacionamentos que permite expressar dados muito complexos (COYLE, 2016, p. 55). No entanto, o modelo de dados em si não faz qualquer suposição sobre o significado dos recursos ou de suas relações. Esse significado fica a cargo dos vocabulários utilizados nas declarações, abordados na seção seguinte.

3.3 *RDF Schema (RDFS) e Web Ontology Language (OWL)*

Com o modelo de dados RDF, abordado na seção anterior, é possível relacionar recursos entre si ou com valores de dados (literais). No entanto, o significado, ou, a semântica, destas relações não é definido pelo modelo de dados em si, mas sim pelas propriedades utilizadas nas declarações. As propriedades, identificadas por URIs, advêm de vocabulários. Nesta pesquisa, emprega-se o entendimento do W3C acerca dos vocabulários:

Na Web Semântica, os vocabulários definem os conceitos e os relacionamentos (também referidos como “termos”) utilizados para descrever e representar uma área de interesse. Os vocabulários são utilizados para classificar os termos usados em uma determinada aplicação, caracterizar os possíveis relacionamentos e definir as possíveis restrições no uso dos termos. Na prática, os vocabulários podem ser muito complexos (com vários milhares de termos) ou muito

simples (descrevendo apenas um ou dois conceitos). (WORLD WIDE WEB CONSORTIUM, 2015, tradução nossa).

Em outras palavras, os vocabulários definem propriedades e/ou classes utilizadas na descrição de recursos. A título de exemplificação, o Quadro 3 lista alguns vocabulários utilizados em declarações RDF e seus respectivos escopos ou domínios de aplicação. Desses vocabulários, aqueles que podem ser utilizados na criação de registros de autoridade são apresentados na Seção 4.1.

Quadro 3 – Exemplos de vocabulários utilizados em declarações RDF

Vocabulário	Escopo e exemplos de propriedades e classes
<p><i>DBpedia</i> http://mappings.dbpedia.org/index.php/How_to_edit_the_DBpedia_Ontology</p>	<p>Descrição de recursos em diversos domínios. Exemplos de propriedades: <i>abbreviation, age, anthem, capital, broadcastArea, publisher</i> e <i>sexualOrientation</i> Exemplos de classes: <i>Activity, Agent, EducationalInstitution, SportsTeam, Person, FictionalCharacter</i> e <i>Monarch</i>.</p>
<p><i>GeoSpecies Ontology</i> http://lod.geospecies.org/ontology/doc/index.html</p>	<p>Descrição de reinos, filos, classes, ordens, famílias, gêneros e espécies biológicas, de registros de ocorrências das espécies e de dados relacionados. Exemplos de propriedades: <i>name, hasCountry, hasStartDate, hasEndDate, hasClassName, hasFamilyName</i> e <i>hasScientificName</i>. Exemplos de classes: <i>Country, Ecozone, Occurrence, Person</i> e <i>Sex</i>.</p>
<p><i>GeoNames Ontology</i> http://www.geonames.org/ontology</p>	<p>Descrição de informações geoespaciais. Exemplos de propriedades: <i>name, officialName, alternateName, population, postalCode, countryCode</i> e <i>wikipediaArticle</i>. Exemplos de classes: <i>Feature, Map</i> e <i>WikipediaArticle</i>.</p>
<p><i>Central Government Ontology</i> http://reference.data.gov.uk/def/central-government</p>	<p>Descrição de dados governamentais do Reino Unido. Exemplos de propriedades: <i>cabinetOf, hasCabinetMember, memberOfCabinet</i> e <i>hasPost</i>. Exemplos de classes: <i>Cabinet, Department, Minister, PublicBody</i> e <i>CivilService</i>.</p>
<p><i>Friend of a Friend (FOAF)</i> http://xmlns.com/foaf/spec/</p>	<p>Descrição de agentes (pessoas, grupos, organizações), suas atividades e relações. Exemplos de propriedades: <i>homepage, knows, accountName, firstName, familyName</i> e <i>gender</i>. Exemplos de classes: <i>Agent, Person, Group, Document, Organization</i> e <i>Image</i>.</p>
<p><i>DCMI Metadata Terms (Dublin Core)</i> http://dublincore.org/documents/dcmi-terms/</p>	<p>Descrição de recursos informacionais. Exemplos de propriedades: <i>title, creator, abstract, subject, identifier, publisher</i> e <i>language</i>. Exemplos de classes: <i>Agent, MediaType, FileFormat, PhysicalMedium</i> e <i>Location</i>.</p>

Vocabulário	Escopo e exemplos de propriedades e classes
<p><i>Simple Knowledge Organization System</i> (SKOS) http://www.w3.org/2004/02/skos/core#</p>	<p>Descrição de sistemas de organização do conhecimento semiformais, tais como tesouros, taxonomias, sistemas de classificação e listas de cabeçalhos de assunto. Exemplos de propriedades: <i>prefLabel, altLabel, broader, narrower, related, scopeNote, exactMatch</i> e <i>closeMatch</i>. Exemplos de classes: <i>Concept, ConceptScheme</i> e <i>Collection</i>.</p>
<p>BIBFRAME http://bibframe.org/</p>	<p>Descrição de recursos informacionais como obras ou instâncias. Exemplos de propriedades: <i>title, content, carrier, originDate, illustrativeContent, classification, edition, extent, dimensions</i> e <i>contributor</i>. Exemplos de classes: <i>Work, Instance, Item, Agent, Person, Organization</i> e <i>Family</i>.</p>
<p><i>RDA Vocabularies</i> http://www.rdaregistry.info/</p>	<p>Descrição de obras, expressões, manifestações, itens, pessoas, famílias e entidades coletivas. Exemplos de propriedades: <i>P30156</i> (título principal), <i>P30133</i> (designação de edição), <i>P30176</i> (nome do publicador), <i>P30011</i> (data de publicação), <i>P20006</i> (idioma da expressão), <i>P20001</i> (tipo de conteúdo), <i>P30001</i> (tipo de suporte), <i>P10223</i> (título preferido para a obra), <i>P50117</i> (nome preferido para a pessoa) e <i>P50121</i> (data de nascimento). Exemplos de classes: <i>C10001</i> (Obra), <i>C10006</i> (Expressão), <i>C10007</i> (Manifestação), <i>C10003</i> (Item), <i>C10002</i> (Agente), <i>C10004</i> (Pessoa), <i>C10008</i> (Família), <i>C10005</i> (Entidade coletiva) e <i>C10009</i> (Lugar).</p>
<p><i>Metadata Authority Description Schema in RDF</i> (MADS/RDF) http://www.loc.gov/standards/mads/rdf/</p>	<p>Descrição de agentes (pessoas, entidades coletivas), eventos e termos (tópicos, geográficos, gêneros, etc.). Exemplos de propriedades: <i>authoritativeLabel, birthDate, associatedLocale, hasAffiliation</i> e <i>hasRelatedAuthority</i>. Exemplos de classes: <i>Authority, Variant, CorporateName, PersonalName</i> e <i>Topic</i>.</p>

Fonte: Elaborado pelo autor.

Além dos vocabulários utilizados para a descrição de recursos de tipos específicos, por exemplo, localizações geográficas ou recursos informacionais do domínio bibliográfico, existem vocabulários que, em termos amplos, são considerados de uso geral. Nesta seção são abordados de forma breve três destes vocabulários que podem ser utilizados na publicação de *Linked Data*: o vocabulário do próprio RDF, o vocabulário do *RDF Schema* (RDFS) e o vocabulário da OWL.⁴²

⁴² Descrições completas desses vocabulários estão disponíveis em suas respectivas especificações, publicadas pela W3C: RDF e *RDF Schema*, disponível em <<https://www.w3.org/TR/rdf-schema/>>, e OWL, disponível em <<https://www.w3.org/TR/2012/REC-owl2-overview-20121211/>>. Acesso em: 12 dez. 2017.

O vocabulário do RDF é identificado pelo URI <http://www.w3.org/1999/02/22-rdf-syntax-ns#>, comumente abreviado como *rdf:*, enquanto o vocabulário do RDFS é identificado por <http://www.w3.org/2000/01/rdf-schema#> e abreviado como *rdfs:* (WORLD WIDE WEB CONSORTIUM, 2014b). Como parte do vocabulário do RDF tem-se a propriedade *rdf:type*, entre outras. Esta propriedade permite declarar que um recurso é uma instância de uma classe, em outras palavras, que um recurso pertence a ou faz parte de uma classe (HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 47-48).

Segundo Coyle (2016, p. 61, tradução nossa)

Estamos todos familiarizados com o conceito de classe dos sistemas de taxonomia científica e de classificação. Nestes sistemas, atribuímos coisas às classes para dar a elas o significado da classe, colocamos os gatos e nós mesmos na classe “mamíferos” e os livros sobre mamíferos em uma das subclasses da biologia.

A autora explica as diferenças entre o significado de classe na Ciência da Informação e na Web Semântica:

As classes têm um significado diferente e funcionam de modo diferente na Web Semântica; elas não são categorias ou caixas para colocar coisas dentro, mas são informações significativas sobre coisas que podem ser utilizadas em vários contextos. As classes não são exclusivas por natureza e qualquer um ou qualquer coisa pode ter qualidades de mais de uma classe. Isso é muito parecido com o mundo real, onde uma pessoa pode ser um empregado em um contexto, um pai em outro e um bombeiro voluntário ainda em outro contexto. Ao invés de atribuir uma coisa a uma classe, a classe é deduzida com base em como algo está descrito. Nossas regras podem dizer que pessoas com contracheques são empregados, que aquelas com filhos são pais, que aquelas que são membros da Brigada Voluntária 7 são bombeiros voluntários, e que qualquer um pode ser todos os três. (COYLE, 2016, p. 61, tradução nossa).

A propriedade *rdf:type*, entendida como “é um(a)”, é utilizada, por exemplo, para declarar que “Machado de Assis é uma Pessoa” e que “Educação é um Conceito”, ou seja, que “Machado de Assis pertence a classe Pessoa” e que “Educação pertence a classe Conceito”. Os valores da propriedade *rdf:type* devem ser recursos, e não valores literais, portanto, devem ser identificados por URIs. Ao ser utilizado como valor da propriedade *rdf:type*, um recurso passa a ser tratado como uma classe, assim, “Pessoa” e “Conceito” são tratados como classes, mesmo que isto não tenha sido declarado em outra tripla. A Figura 15 apresenta estes dois exemplos serializados no formato *Turtle*.

Figura 15 – Declarações utilizando a propriedade *rdf:type* serializadas em *Turtle*.

"Machado de Assis é uma Pessoa"

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix vocabp: <http://vocab.org/pessoas#> .

<http://pessoas.com/entidade/12345>
  rdf:type <http://vocab.org/classes/Pessoa> ;
  vocabp:nome "Machado de Assis" ;
  vocabp:data-de-nascimento "1839-06-21" ;
  vocabp:conjuge <http://pessoas.com/entidade/67890> .
```

"Educação é um Conceito"

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix tesau: <http://propriedades.org/tesauoro#> .

<http://tesauoro.com.br/conceitos/0026>
  rdf:type <http://vocab.org/classes/Conceito> ;
  tesau:termo-autorizado "Educação"@pt ;
  tesau:termo-autorizado "Education"@en ;
  tesau:termo-autorizado "Educación"@es .
```

Fonte: Elaborada pelo autor.

Além do vocabulário do RDF, é comum o uso do vocabulário do RDFS, que, enquanto “[...] uma extensão do RDF, proporciona um meio para a publicação tanto de vocabulários legíveis por humanos quanto de vocabulários legíveis por máquinas” (FERREIRA; SANTOS, 2013, p. 15). Para Hitzler, Krötzsch e Rudolph (2010, p. 46), o RDFS permite especificar o conhecimento terminológico (*terminological knowledge*) acerca dos termos utilizados em um vocabulário.

Por meio do RDFS, é possível, por exemplo:

- declarar rótulos (“nomes”) legíveis por humanos para recursos, propriedades e classes, por exemplo, o rótulo “data de nascimento” para a propriedade com o URI *http://vocab.org/pessoas#data-de-nascimento* e o rótulo “Pessoa” para a classe *http://vocab.org/classes/Pessoa*; para a declaração destes rótulos é utilizada a propriedade *rdfs:label*, que tem como valor um valor literal;
- declarar que um recurso é uma classe, por exemplo, declarar que *http://vocab.org/classes/Pessoa* é uma classe; para esta declaração é utilizada a propriedade *rdf:type* como o valor *rdfs:Class*;

- declarar a quais classes uma determinada propriedade se aplica, por exemplo, declarar que a propriedade “data de fundação” se aplica aos recursos da classe “Entidade coletiva” e que a propriedade “cônjuge” se aplica à classe “Pessoa”; para esta declaração é utilizada a propriedade *rdfs:domain*;⁴³
- declarar quais valores são esperados para uma determinada propriedade, por exemplo, declarar que a propriedade “cônjuge” terá como valor um recurso da classe “Pessoa” e que a propriedade “termo” terá como valor um valor literal; para esta declaração é utilizada a propriedade *rdfs:range*;⁴³
- declarar as relações hierárquicas existentes entre propriedades e classes, por exemplo, declarar que as classes “Conferência”, “Órgão governamental” e “Grupo musical” são subclasses da classe “Entidade coletiva” e que as propriedades “termo autorizado” e “termo não autorizado” são subpropriedades da propriedade “termo”; para estas declarações são utilizadas as propriedades *rdfs:subClassOf* e *rdfs:subPropertyOf*, respectivamente (WORLD WIDE WEB CONSORTIUM, 2014b; HITZLER; KRÖTZSCH; RUDOLPH, 2010).

Estes exemplos (*rdfs:label*, *rdfs:Class*, *rdfs:domain*, *rdfs:range*, *rdfs:subClassOf* e *rdfs:subPropertyOf*), apresentados em *Turtle* na Figura 16, são apenas algumas das propriedades e classes do RDFS.⁴⁴

Figura 16 – Declarações utilizando propriedades e classes do *RDF Schema*

Propriedade *rdfs:label*

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://vocab.org/pessoas#data-de-nascimento>
  rdfs:label "Data de nascimento"@pt ;
```

⁴³ Uma vez que o RDF tem em sua base a realização de inferências sobre os dados e não o controle sobre eles, as propriedades *rdfs:domain* e *rdfs:range* não são utilizadas para a restrição do domínio e do valor de uma propriedade, mas sim para a realização de inferências sobre o recurso e o valor de uma declaração. Exemplo: se a propriedade “isbn” for declarada tendo como *rdfs:domain* a classe “Livro”, um recurso que utilizar esta propriedade em sua descrição será considerado um indivíduo da classe “Livro”, mesmo que, em outra tripla, tenha sido declarado como indivíduo da classe “Artigo de periódico”. Ao processar este exemplo, um motor de inferência não identificará qualquer inconsistência, uma vez que um recurso pode pertencer a mais de uma classe. No entanto, se em outra tripla for declarado que as classes “Livro” e “Artigo de periódico” são disjuntas, ou seja, que um recurso não pode pertencer a essas duas classes ao mesmo tempo, o motor de inferência identificará uma inconsistência. A declaração de classes disjuntas é possibilitada pela OWL (BAKER; COYLE; PETIYA, 2014, p. 567; HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 55-56).

⁴⁴ Uma lista completa está disponível na especificação publicada pelo W3C (WORLD WIDE WEB CONSORTIUM, 2014b).

rdfs:label "Date of birth"@en .

<http://vocab.org/classes#Pessoa> rdfs:label "Pessoa" .

Propriedade *rdfs:type* e classe *rdfs:Class*

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://vocab.org/classes#EntidadeColetiva>
 rdfs:type rdfs:Class ;
 rdfs:label "Entidade coletiva" .

Propriedades *rdfs:domain* e *rdfs:range*

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://vocab.org/grupos#fundacao>
 rdfs:label "Data de fundação" ;
 rdfs:domain <http://vocab.org/classes#EntidadeColetiva> .

<http://vocab.org/pessoas#conjuge>
 rdfs:label "Cônjuge" ;
 rdfs:domain <http://vocab.org/classes#Pessoa> ;
 rdfs:range <http://vocab.org/classes#Pessoa> .

<http://propriedades.org/tesauro#termo>
 rdfs:label "Termo" ;
 rdfs:range rdfs:Literal .

Propriedade *rdfs:subClassOf*

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://vocab.org/classes#Conferencia>
 rdfs:label "Conferência" ;
 rdfs:subClassOf <http://vocab.org/classes#EntidadeColetiva> .

<http://vocab.org/classes#Banda>
 rdfs:label "Grupo musical" ;
 rdfs:subClassOf <http://vocab.org/classes#EntidadeColetiva> .

<http://vocab.org/classes#OrgaoGovernamental>
 rdfs:label "Órgão governamental" ;
 rdfs:subClassOf <http://vocab.org/classes#EntidadeColetiva> .

Propriedade *rdfs:subPropertyOf*

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://propriedades.org/tesauro#termo-autorizado>
 rdfs:label "Termo autorizado" ;
 rdfs:subPropertyOf <http://propriedades.org/soc#termo> .

<http://propriedades.org/tesauro#termo-nao-autorizado>
 rdfs:label "Termo não autorizado" ;
 rdfs:subPropertyOf <http://propriedades.org/soc#termo> .

Fonte: Elaborada pelo autor.

O RDF, como destacado por Ferreira e Santos (2013, p. 21),

[...] oferece a possibilidade para as comunidades de descrição de recursos definirem a semântica de seus metadados de maneira formal, isto é, definindo o significado dos elementos de metadados, conforme as suas necessidades específicas de descrição, em um modelo processável por máquinas.

Determinadas comunidades (ou domínios) podem apresentar necessidades de descrição que demandem relacionamentos entre recursos mais complexos que aqueles possibilitados pelas classes e propriedades do RDF e do RDFS. Tais comunidades podem encontrar no vocabulário da *Web Ontology Language* (OWL) (WORLD WIDE WEB CONSORTIUM, 2012a) classes e propriedades que atendam as suas demandas.

Publicada primeiramente em 2004, e hoje em sua segunda edição (publicada em 2012), a OWL é definida como

[...] uma linguagem de ontologia para a Web Semântica com significado formalmente definido. As ontologias com a OWL 2 proveem classes, propriedades, indivíduos e valores de dados e são armazenadas como documentos da Web Semântica. As ontologias com a OWL 2 podem ser utilizadas com informações escritas em RDF e são principalmente intercambiadas como documentos RDF. (WORLD WIDE WEB CONSORTIUM, 2012a, tradução nossa).

Para Hitzler, Krötzsch e Rudolph (2010, p. 2, tradução nossa), “[...] na ciência da computação, uma ontologia é uma descrição do conhecimento sobre um domínio de interesse, o cerne da qual é uma especificação processável por máquina com um significado formalmente definido”. Contextualizando as ontologias na Ciência da Informação, Ramalho (2010, p. 107) define ontologia como

um artefato tecnológico que possibilita representar formalmente as propriedades e relacionamentos de um determinado modelo conceitual, favorecendo a utilização de inferências automáticas nos processos de organização e recuperação de recursos informacionais.

Baker, Coyle e Petiya (2014, p. 565, tradução nossa), entendem que “vocabulários em RDF definidos com muitos axiomas específicos são às vezes chamados de ontologias”. Para o W3C não há uma distinção clara entre vocabulários e ontologias, sendo uma tendência o uso de “ontologias” para os conjuntos de termos mais complexos e com formalismo rigoroso, e o uso de “vocabulários” para os conjuntos em que tal

formalismo não é necessariamente utilizado ou é pouco utilizado (WORLD WIDE WEB CONSORTIUM, 2015).

Willer e Dunsire (2013, p. 112, tradução nossa) apresentam um conceito de ontologia e o relaciona aos conceitos de RDF e RDFS:

Uma ontologia é uma representação formal da estrutura de conhecimento e informação. Os grafos RDF que descrevem somente coisas que são classes e propriedades são ontologias. As propriedades destas classes e propriedades representam os relacionamentos semânticos de um determinado sistema de conhecimento.

Ontologias que são complexas necessitam de propriedades ontológicas que não estão disponíveis no *namespace* do RDF e do RDFS. Como uma alternativa, elas estão descritas na Web Ontology Language (OWL), que possui seu próprio *namespace*. (WILLER; DUNSIRE, 2013, p. 113, tradução nossa).

Para Heath e Bizer (2011, seção 4.4.1, tradução nossa), o RDFS é frequentemente suficiente para a publicação de *Linked Data*, no entanto, alguns casos que demandam uma maior expressividade podem requerer o uso de recursos da OWL:

RDFS e OWL são utilizados nos casos em que a subordinação de relacionamentos entre termos deve ser representada (por exemplo, que todos os atletas são também pessoas). Quando combinados com um motor de inferência apropriado, os modelos RDFS e OWL permitem que relacionamentos implícitos sejam inferidos aos dados. Em um contexto de *Linked Data*, geralmente é suficiente expressar vocabulários em RDFS. No entanto, certas primitivas da OWL, tais como *sameAs*, são utilizadas regularmente para declarar que dois URIs identificam o mesmo recurso.

Assim como o RDFS expande o RDF (FERREIRA; SANTOS, 2013, p. 15), pode-se dizer também que a OWL expande o RDFS:

A OWL estende a expressividade do RDFS com primitivas adicionais de modelagem. Por exemplo, a OWL define as primitivas *owl:equivalentClass* e *owl:equivalentProperty*. Quando combinadas com *rdfs:subClassOf* e *rdfs:subPropertyOf*, elas proveem poderosos mecanismos para a definição de mapeamentos entre termos de diferentes vocabulários, o que, por sua vez, aumenta a interoperabilidade dos conjuntos de dados modelados utilizando diferentes vocabulários [...] (HEATH; BIZER, 2011, seção 4.4.3, tradução nossa).

Algumas das propriedades da OWL utilizadas na publicação de *Linked Data* possibilitam:⁴⁵

⁴⁵ Uma apresentação mais detalhada da OWL e de suas propriedades é realizada por Santarem Segundo e Coneglian (2016).

- declarar que recursos identificados por diferentes URIs representam a mesma entidade, por exemplo, declarar que `http://pessoas.com/entidade/12345` e `http://www.wikidata.org/entity/Q311145` identificam a mesma pessoa; para essa declaração é utilizada a propriedade `owl:sameAs`;
- declarar que propriedades identificadas por diferentes URIs são equivalentes (expressam o mesmo tipo de relação), por exemplo, declarar que a propriedade `http://vocab.org/pessoas#conjuge` tem como equivalente a propriedade `http://dbpedia.org/ontology/spouse` da DBpedia; para essa declaração é utilizada a propriedade `owl:equivalentProperty`;
- declarar que classes identificadas por diferentes URIs são equivalentes, por exemplo, declarar que a classe `http://vocab.org/classes#Pessoa` tem com classe equivalente a classe `http://xmlns.com/foaf/0.1/Person` do FOAF; para essa declaração é utilizada a propriedade `owl:equivalentClass` (HITZLER; KRÖTZSCH; RUDOLPH, 2010, p. 118, 120, 133; WORLD WIDE WEB CONSORTIUM, 2012b).

Exemplos de uso destas propriedades são apresentados na Figura 17 por meio de declarações em RDF serializadas em *Turtle*.

Figura 17 – Declarações em RDF utilizando algumas propriedades da OWL

Propriedade `owl:sameAs`

```
@prefix vocabp: <http://vocab.org/pessoas#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://pessoas.com/entidade/12345>
  vocabp:nome "Machado de Assis" ;
  owl:sameAs <http://www.wikidata.org/entity/Q311145> .
```

Propriedade `owl:equivalentProperty`

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://vocab.org/pessoas#conjuge>
  rdfs:label "Cônjuge" ;
  owl:equivalentProperty <http://dbpedia.org/ontology/spouse> .
```

Propriedade `owl:equivalentClass`

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://vocab.org/classes#Pessoa>
  rdfs:label "Pessoa" ;
```

owl:equivalentClass <http://xmlns.com/foaf/0.1/Person> .

Fonte: Elaborada pelo autor.

A OWL, ao possibilitar a declaração de equivalências entre recursos, propriedades e classes, contribui de forma significativa para a quarta e última das melhores práticas para a publicação de *Linked Data*: “Inclua *links* para outros URIs” (BERNERS-LEE, 2006, tradução nossa).

As propriedades da OWL permitem contornar ou, ao menos, amenizar, as dificuldades para o uso e a gestão dos dados decorrentes da multiplicidade de URIs utilizados na identificação de um mesmo recurso, sendo essa multiplicidade parte da premissa do nome não-único (*Non-Unique Name Assumption* – NUNA), abordada na Seção 3.2.

Os *links* declarados entre recursos de diferentes conjuntos de dados são fundamentais, “[...] uma vez que são a cola que conecta as ilhas de dados em um espaço de dados global interconectado e que permite às aplicações descobrirem fontes de dados adicionais [...]” (HEATH; BIZER, 2011), auxiliando, assim, na integração de informações e na criação de novos caminhos para a descoberta de outras informações (PATTUELLI, 2012, p. 558).

Fundamentais também são os links estabelecidos entre propriedades e classes de diferentes vocabulários. Estes links, possibilitados pela OWL, permitem o mapeamento de vocabulários de um mesmo domínio ou de domínios diversos, o que favorece a reutilização não só das descrições dos recursos, mas também das propriedades e classes.

As melhores práticas para a publicação de *Linked Data* incluem também o uso de *SPARQL Protocol and RDF Query Language* (SPARQL) para a consulta aos dados. Segundo Santarem Segundo e Coneglian (2016, p. 224),

O SPARQL é um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web. O padrão compreende as seguintes especificações: uma linguagem de consulta para RDF; uma especificação que define uma extensão do *SPARQL Query Language* para executar consultas distribuídas em diferentes terminais SPARQL; uma especificação que define a semântica de consultas SPARQL sob regimes de vinculação, como *RDF Schema*, OWL, ou RIF; um protocolo que define os meios para a transmissão de consultas SPARQL arbitrárias e solicitações de atualização para um serviço de SPARQL; uma especificação que define um método de busca e descoberta e um vocabulário para descrever serviços SPARQL e um conjunto de testes, para avaliação da especificação SPARQL 1.1.

A opção por não abordar tal tecnologia de forma detalhada decorre do entendimento de que o SPARQL relaciona-se mais ao desenvolvimento, por profissionais da Ciência da Computação, de aplicações tecnológicas para o acesso aos dados do que ao papel dos profissionais da Ciência da Informação na publicação de *Linked Data*.⁴⁶ No entanto, mesmo sem receber uma abordagem detalhada, o SPARQL será retomado mais adiante, na Seção 5.4.2, em aborda a publicação de dados em formatos acessíveis aos softwares.

No âmbito do *Linked Data*, as tecnologias abordadas neste capítulo (URIs, RDF, RDFS e OWL) prestam-se principalmente à publicação de dois tipos de dados: (1) conjuntos de dados com descrições de recursos e (2) vocabulários com descrições de propriedades e de classes. Após a apresentação destas tecnologias e dos conceitos a elas relacionados, no capítulo seguinte são descritas iniciativas para a publicação de conjuntos de dados e de vocabulários no domínio bibliográfico com foco nos dados de autoridade, descrição esta essencial ao objetivo geral desta pesquisa.

⁴⁶ Uma apresentação detalhada e didática do SPARQL é conduzida por Santarem Segundo (2014).

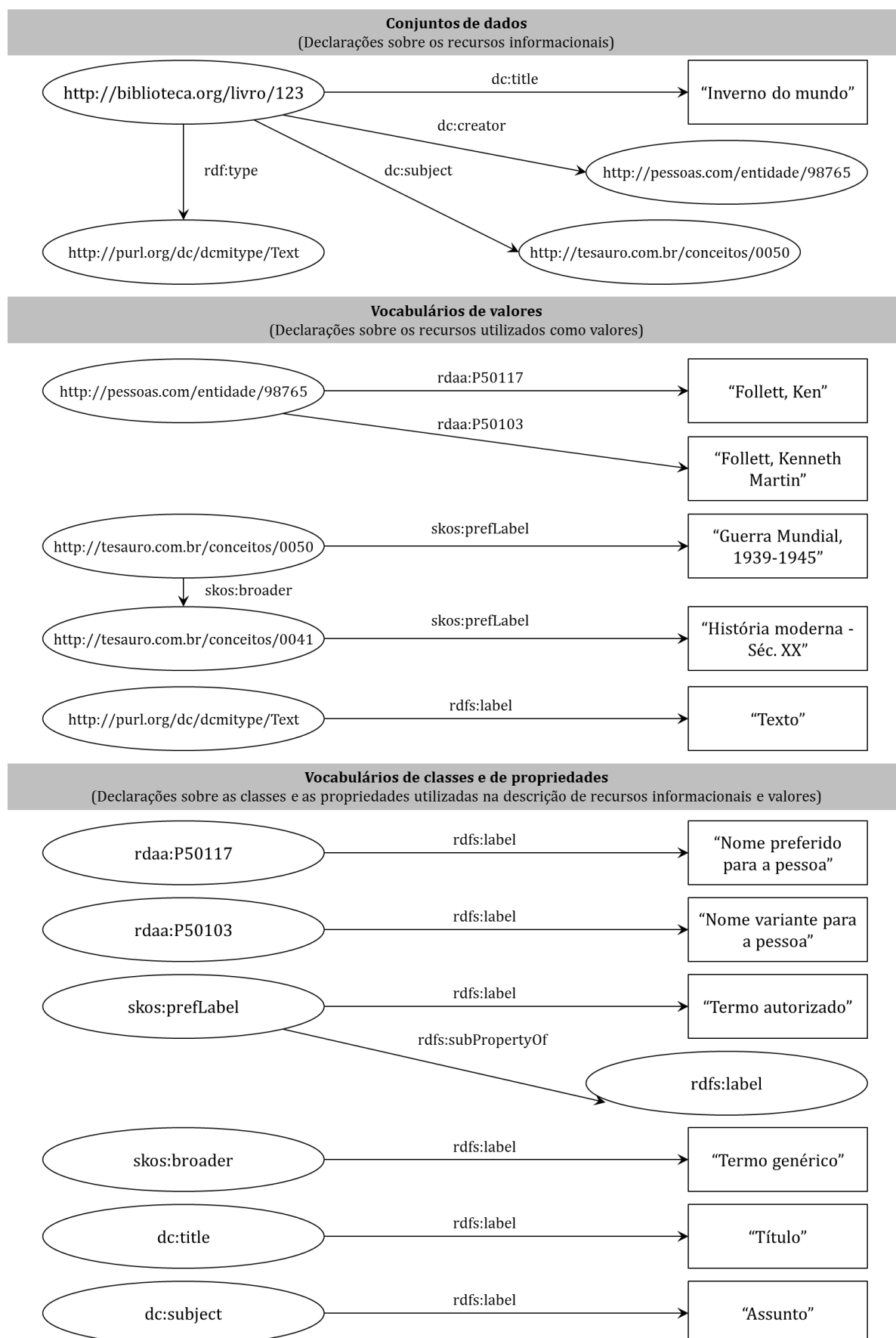
4 *Linked Data* no domínio bibliográfico: os dados de autoridade em foco

No capítulo anterior foram apresentados o conceito de *Linked Data* e algumas tecnologias da Web Semântica a ele relacionadas. Utilizando-se desse conceito e dessas tecnologias, conjuntos de dados com descrições de recursos e vocabulários de classes e de propriedades têm sido publicados como *Linked Data* em diversos domínios, inclusive no domínio bibliográfico.

Como importante iniciativa para a utilização das tecnologias da Web Semântica no domínio bibliográfico, destaca-se a criação, em 2010, do *Library Linked Data Incubator Group* pelo W3C “para auxiliar no aumento global da interoperabilidade dos dados das bibliotecas na Web” (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2012, tradução nossa). Em seu relatório, o grupo apresenta como principais recomendações:

- Que os **líderes de bibliotecas** identifiquem conjuntos de dados como possíveis candidatos para uma exposição inicial como *Linked Data* e fomentem uma discussão sobre *Open Data* e direitos;
- Que as **entidades de padronização das bibliotecas** aumentem a participação das bibliotecas na padronização da Web Semântica, desenvolvam padrões para os dados de bibliotecas compatíveis com o *Linked Data* e disseminem as melhores práticas adaptadas para o *Linked Data* em bibliotecas;
- Que os **projetistas de dados e de sistemas** projetem serviços melhorados aos usuários baseados nas capacidades do *Linked Data*, criem URIs para os itens dos conjuntos de dados das bibliotecas, desenvolvam políticas para a gestão de vocabulários em RDF e seus URIs, e expressem os dados das bibliotecas reutilizando os vocabulários de *Linked Data* existentes ou mapeando para eles;
- Que os **bibliotecários e os arquivistas** preservem conjuntos de elementos e vocabulários de valores do *Linked Data* e apliquem a experiência das bibliotecas na curadoria e na preservação de longo prazo aos conjuntos de dados do *Linked Data*. (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, grifo do autor, tradução nossa).

Figura 18 – Categorias de dados das bibliotecas



Fonte: Elaborada pelo autor.

No relatório são definidas três categorias, não exclusivas, de dados das bibliotecas, exemplificadas na Figura 18:

- **conjuntos de dados:** conjuntos de descrições de recursos, tais como os livros de uma biblioteca;
- **vocabulários de valores:** conjuntos de recursos utilizados como valores ou, ainda, listas de valores controlados para o uso em determinados elementos, por exemplo, arquivos de autoridade de nomes, tesouros, listas de cabeçalhos de assunto, sistemas de classificação, taxonomias e listas de códigos;
- **conjuntos de elementos** (também denominados conjuntos de elementos de metadados ou vocabulários de classes e de propriedades⁴⁷): conjuntos de classes e de propriedades utilizadas para descrever as entidades de interesse (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011; WILLER; DUNSIRE, 2013, p. 115)

Vale reiterar que essa distinção entre conjuntos de dados, vocabulários de valores e vocabulários de classes e de propriedades baseia-se no uso dos dados no âmbito das bibliotecas e que, do ponto de vista do modelo RDF, todo recurso pode ser em algum momento utilizado como valor de uma declaração, e uma propriedade ou uma classe, embora desempenhe tal função em uma declaração, em outra declaração pode exercer também a função de recurso ou de valor.

Uma vez que essa pesquisa volta-se aos dados de autoridade, são de seu interesse apenas os vocabulários de valores, compostos por dados de autoridade, e os vocabulários de classes e de propriedades utilizados na descrição dos dados de autoridade. Os conjuntos de dados, enquanto conjuntos de descrições de recursos informacionais, não são abordados aqui, pois fogem do escopo dessa pesquisa, já que tratam de dados bibliográficos e não de dados de autoridade.

A abordagem das categorias selecionadas ocorre em duas seções: Seção 4.1, em que são tratados vocabulários de classes e de propriedades, expressos em RDF, utilizados na descrição de dados de autoridade, são eles SKOS, MADS, *RDA Element Sets*; e Seção 4.2, que trata de iniciativas para a publicação de dados de autoridade como *Linked Data*.

⁴⁷ Esse último termo, vocabulários de classes e de propriedades, embora não conste no relatório, é utilizado nesta pesquisa como sinônimo de conjuntos de elementos. A escolha por esse termo deu-se para a manutenção da conformidade com os termos empregados no Capítulo 3.

4.1 Vocabulários para a descrição de dados de autoridade⁴⁸

Como abordado na Seção 3.3, a semântica da relação entre recursos ou entre um recurso e um valor literal é dada pela propriedade utilizada na declaração. As relações entre as propriedades são declaradas em vocabulários. Para a Web Semântica é necessário o uso de vocabulários de classes e de propriedades, às vezes denominados ontologias, cuja semântica possa ser processada por máquinas, e não apenas por humanos.

Para Santarem Segundo (2015, p. 226), a publicação de *Linked Data* com vistas a uma recuperação da informação eficiente requer não somente a publicação de declarações em RDF sobre os recursos descritos, mas, também, o uso de ontologias e de vocabulários internacionalmente reconhecidos:

[...] para que se possa publicar dados e principalmente torná-los disponíveis para serem recuperados com mais eficiência e eficácia, é necessário que o esquema lógico esteja sob uma ontologia, e de preferência utilizando-se de vocabulários de representação de conhecimento padronizados e reconhecidos internacionalmente. (SANTAREM SEGUNDO, 2015, p. 226).

Sobre o não uso de ontologias ou de vocabulários já reconhecidos, o autor complementa que

Não utilizar esquemas lógicos como as ontologias e os principais vocabulários para descrever dados a serem publicados certamente é o pior caminho para publicação de dados em formato semântico, principalmente do ponto de vista da recuperação da informação. (SANTAREM SEGUNDO, 2015, p. 226).

Pattuelli (2012, p. 559) destaca que os vocabulários “[...] são ferramentas chave para facilitar a integração e o reuso de conteúdo, dada sua capacidade de reduzir a ambiguidade semântica e de sustentar efetivamente os relacionamentos entre diferentes conjuntos de dados”.

Os vocabulários de classes e de propriedades assemelham-se aos formatos ou padrões de metadados tradicionalmente tratados na Ciência da Informação. Por padrões de metadados, entende-se aqui “[...] estruturas de descrição constituídas por um conjunto predeterminado de metadados (atributos codificados ou identificadores de

⁴⁸ O conteúdo desta seção está parcialmente publicado no trabalho “*Linked Data* no domínio bibliográfico: vocabulários para a publicação de dados de autoridade” (ASSUMPÇÃO; SANTOS, 2016), apresentado no 3º Encontro Internacional de Dados, Tecnologia e Informação (DTI).

uma entidade) metodologicamente construídos e padronizados” (ALVES, 2010, p. 47-48). Ao relacionar o conceito de vocabulário, utilizado na Web Semântica, ao conceito de padrão de metadados, utilizado na Ciência da Informação, é necessário considerar que todo padrão de metadados é um vocabulário, mas que nem todo padrão de metadados é um vocabulário apto para ser utilizado na Web Semântica. Além disso, a Ciência da Informação, em razão de seus objetos de estudo, tem restringido seu conceito de padrão de metadados à representação de recursos informacionais, enquanto que o conceito de vocabulário na Web Semântica volta-se aos recursos de qualquer natureza, informacionais ou não.

Nesta seção são apresentados três vocabulários que podem ser utilizados na publicação de dados de autoridade como Linked Data: o *Simple Knowledge Organization System* (SKOS) (Seção 4.1.1), o *Metadata Authority Description Schema in RDF* (MADS/RDF) (Seção 4.1.2) e os *RDA Element Sets* (Seção 4.1.3).

4.1.1 *Simple Knowledge Organization System* (SKOS)

O *Simple Knowledge Organization System* (SKOS) (Sistema simples de organização do conhecimento) é “[...] um vocabulário RDF para a representação de sistemas de organização do conhecimento semi-formais, tais como tesouros, taxonomias, sistemas de classificação e listas de cabeçalhos de assunto” (WORLD WIDE WEB CONSORTIUM, 2009, tradução nossa). O SKOS foi projetado para permitir, com um baixo custo, que sistemas de organização do conhecimento existentes sejam migrados para a Web Semântica, ou seja, expressos utilizando tecnologias da Web Semântica como o RDF (WORLD WIDE WEB CONSORTIUM, 2009).

Sobre o desenvolvimento do SKOS, Pastor Sánchez, Martínez Méndez e Rodríguez Muñoz (2012, p. 246, tradução nossa) sintetizam que ele

[...] teve início no grupo de trabalho SWAD-Europe [*Semantic Web Advanced Development for Europe*] até 2002 e difundiu-se publicamente em um rascunho em novembro de 2005. Naquele momento, a proposta fez-se conhecer como *SKOS Core*, nome que ainda perdura em muitos trabalhos atuais. Em agosto de 2009, o SKOS alcançou o status de uma recomendação do W3C.

No SKOS, os sistemas de organização do conhecimento são tratados como esquemas de conceitos.⁴⁹ Para os conceitos que integram um esquema são declarados rótulos, assim como relações com outros conceitos do próprio esquema e com conceitos de outros esquemas (WORLD WIDE WEB CONSORTIUM, 2009). Algumas das classes e das propriedades do SKOS utilizadas nessas declarações estão exemplificadas na Figura 19.

Figura 19 – Exemplos de classes e propriedades do SKOS

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://tesouro.com.br/conceitos/0031>
  rdf:type skos:Concept ;
  skos:prefLabel "Animais domésticos"@pt ;
  skos:prefLabel "Domestic animals"@en ;
  skos:altLabel "Animais de criação"@pt ;
  skos:broader <http://tesouro.com.br/conceitos/0028> ;
  skos:narrower <http://tesouro.com.br/conceitos/0035> ;
  skos:inScheme <http://tesouro.com.br/conceitos> .

<http://tesouro.com.br/conceitos/0035>
  rdf:type skos:Concept ;
  skos:prefLabel "Animais de estimação"@pt ;
  skos:scopeNote "Usado para animais domésticos selecionados para o convívio com os seres humanos por questões de companheirismo ou de divertimento" ;
  skos:inScheme <http://tesouro.com.br/conceitos> .

<http://tesouro.com.br/conceitos/0028>
  rdf:type skos:Concept ;
  skos:prefLabel "Animais"@pt ;
  skos:related <http://tesouro.com.br/conceitos/0010> ;
  skos:inScheme <http://tesouro.com.br/conceitos> .

<http://tesouro.com.br/conceitos/0010>
  rdf:type skos:Concept ;
  skos:prefLabel "Zoologia"@pt ;
  skos:inScheme <http://tesouro.com.br/conceitos> ;
  skos:exactMatch <http://id.loc.gov/authorities/subjects/sh85149983> .

<http://tesouro.com.br/conceitos>
  rdf:type skos:ConceptScheme ;
  dc:title "Tesauro para bibliotecas públicas"@pt .

```

Fonte: Elaborada pelo autor.

⁴⁹ Van Hooland e Verborgh (2015, p. 130, tradução nossa) chamam a atenção para distinção entre conceito e termo: “É importante compreender que os conceitos são a base do SKOS, não os termos. Essa distinção também está incorporada nas novas diretrizes da ISO para tesauros. Conceitos podem ser abstratos e podem ser expressos de diferentes maneiras a partir de termos concretos”.

As classes e as propriedades do SKOS exemplificadas na Figura 19 possibilitam, entre outros:

- declarar que um recurso é um conceito (pertence à classe *skos:Concept*);
- declarar rótulos preferidos (pontos de acesso/termos autorizados) (*skos:prefLabel*) e alternativos (pontos de acesso/termos não autorizados) (*skos:altLabel*) para os conceitos;
- declarar relações hierárquicas entre os conceitos indicando conceitos mais genéricos (*skos:broader*) e conceitos mais específicos (*skos:narrower*);
- declarar relações associativas entre os conceitos (*skos:related*);
- declarar notas relacionadas ao conceito: nota de definição (*skos:definition*), nota de escopo (*skos:scopeNote*), nota de exemplo (*skos:example*), nota histórica (*skos:historyNote*), nota editorial (*skos:editorialNote*) e nota de alteração (*skos:changeNote*);
- declarar que um recurso é um esquema de conceitos (pertence à classe *skos:ConceptScheme*);
- declarar a quais esquemas de conceitos um conceito está relacionado (*skos:inScheme*);
- declarar relações entre conceitos de diferentes esquemas: correspondência exata (*skos:exactMatch*), correspondência aproximada (*skos:closeMatch*), correspondência com conceito mais genérico (*skos:broadMatch*), correspondência com conceito mais específico (*skos:narrowMatch*) e correspondência com conceito relacionado (*skos:relatedMatch*).

Além das relações entre conceitos, são declaradas também relações entre os rótulos com o uso do *SKOS extension for Labels* (SKOS-XL) (PASTOR SÁNCHEZ; MARTÍNEZ MÉNDEZ; RODRIGUÉZ MUÑOZ, 2012, p. 246).

Sobre a aceitação e a aplicabilidade do SKOS para a representação de sistemas de organização do conhecimento, Catarino (2014, p. 26) aponta que

[...] o SKOS tem sido aceito pela comunidade responsável pela organização [da] informação, tendo em vista que vários vocabulários controlados utilizados em todo o mundo, já estão convertidos para o modelo. Portanto, trata-se de um modelo que deve ser aplicado na construção de esquemas de conceito que venham a contribuir para a proposta de Dados Lincados do W3C visando a Web Semântica.

A aceitação e a aplicabilidade às quais a autora refere-se são observadas no trabalho de Pastor Sánchez, Martínez Méndez e Rodríguez Muñoz (2012). A partir de dados obtidos na plataforma *Datahub*⁵⁰ e em fontes complementares, os autores localizaram 55 vocabulários controlados que utilizam o SKOS; desses vocabulários foram identificados, entre outras variáveis, o tipo de vocabulário controlado e a quantidade de declarações (triplas), apresentados na Tabela 1.

Tabela 1 – Vocabulários que utilizam o SKOS (tipos e quantidade de declarações)

Tipo de vocabulário	Vocabulários analisados	Frequência (%)	Quantidade de triplas	Volume de dados (%)
Tesauros	19	34,6	28.309.759	8,2
Classificações	16	29,1	6.891.148	2,0
Cabeçalhos de assunto	7	12,7	36.724.495	10,6
Arquivos de autoridade	4	7,3	255.528.450	73,6
Léxicos e glossários	4	7,3	388.729	0,1
Ontologias	3	5,5	5.003.228	1,4
Taxonomias	2	3,6	14.542.505	4,2
Totais	55	100	347.388.314	100

Fonte: Pastor Sánchez, Martínez Méndez e Rodríguez Muñoz (2012, p. 248, tradução nossa).

Pelos dados da Tabela 1, observa-se que o SKOS tem sido utilizado de forma relevante na publicação de arquivos de autoridade de nomes, que não são considerados sistemas de organização do conhecimento. Ainda que o uso do SKOS tenha sido identificado em apenas quatro arquivos de autoridade, esses quatro representam 73,6 % do volume total de dados (quantidade de triplas) dos 55 vocabulários controlados localizados pelos autores. Para van Hooland e Verborgh (2015, p. 128), o SKOS é, de longe, o mais bem sucedido entre os esforços para a padronização dos vocabulários.

Sobre a aplicabilidade do SKOS, Ramalho (2016, p. 8), aponta que,

Podendo ser utilizado de forma isolada ou em combinação com linguagens formais de representação, como OWL (*Ontology Web Language*), SKOS fornece um elo entre o formalismo lógico rigoroso das linguagens de ontologias e o mundo caótico, informal e pouco estruturado de ferramentas de colaboração baseadas na Web.

⁵⁰ Disponível em: <<https://datahub.io/>>.

A simplicidade do SKOS, junto de sua capacidade de representar não somente sistemas de organização do conhecimento, mas qualquer recurso ao qual possam ser aplicados, por exemplo, rótulos preferidos e alternativos, tem contribuído para o seu uso em diversos domínios ou comunidades. No entanto, em alguns casos, “[...] ele não alcança o nível de detalhamento dos dados de autoridade e de vocabulário da Biblioteconomia e da Ciência da Informação” (LIBRARY OF CONGRESS, 2015, tradução nossa). Na tentativa de atender a esse nível de detalhamento, outros vocabulários têm sido desenvolvidos no âmbito da Ciência da Informação, alguns dos quais são apresentados nas subseções seguintes.

4.1.2 *Metadata Authority Description Schema in RDF (MADS/RDF)*

O *Metadata Authority Description Schema* (MADS) (Esquema de metadados para a descrição de autoridade) teve sua primeira versão publicada em 2005 como um acompanhamento para o *Metadata Object Description Schema* (MODS) (Esquema de metadados para a descrição de objeto), criado em 2002. Os dois padrões de metadados foram criados pela LC como alternativas à utilização dos Formatos MARC 21 para Dados Bibliográficos e para Dados de Autoridade: o MODS seria utilizado com registros bibliográficos descrevendo recursos informacionais, enquanto que o MADS seria utilizado em registros de autoridade descrevendo os pontos de acesso controlados utilizados nos registros bibliográficos (ASSUMPÇÃO; SANTOS, 2013b).

Inicialmente, o MADS foi publicado como um esquema XML⁵¹ com o objetivo de definir uma estrutura para a publicação de registros de autoridade de agentes (pessoas, organizações), de eventos e de termos (tópicos, geográficos, de gêneros, etc.) como documentos XML. Essa estrutura, além de contemplar os elementos permitidos em um documento XML, apresenta também a hierarquia dos elementos dentro do documento e as restrições aplicadas aos seus conteúdos.

A partir do MADS foi criado o *Metadata Authority Description Schema in RDF* (MADS/RDF), definido como “[...] um modelo de dados para dados de autoridade e de vocabulário usado na Biblioteconomia e na Ciência da Informação, que inclui também

⁵¹ Um esquema XML é um documento criado com a linguagem *XML Schema* com o objetivo de definir a estrutura, os elementos permitidos, etc. de um determinado tipo ou conjunto de documentos XML (ASSUMPÇÃO; SANTOS, 2013b, p. 113).

museus, arquivos e outras instituições culturais” (LIBRARY OF CONGRESS, 2015, tradução nossa).

As principais metas que guiam o desenvolvimento do MADS/RDF são:

- interoperabilidade com o SKOS: as classes e propriedades do MADS/RDF estão mapeadas como subclasses e subpropriedades das classes e propriedades do SKOS, por exemplo, a classe *madsrdf:Authority* é uma subclasse de *skos:Concept* e a propriedade *madsrdf:authoritativeLabel* é uma subpropriedade de *skos:prefLabel*;
- suportar os tipos de registros de autoridade utilizados nos sistemas de gestão de bibliotecas: é possível indicar o tipo de registro de autoridade, por exemplo, nome pessoal, nome de entidade coletiva, título, área geográfica, conceito temporal, etc.;
- utilizar o MADS/XML como base para o MADS/RDF: os dados descritos com o MADS/XML podem ser descritos também com o MADS/RDF;
- suportar descrições e rótulos autorizados, variantes e obsoletos: é possível indicar quais rótulos são autorizados e não autorizados e quais estão obsoletos;
- suportar partes de rótulos: os rótulos podem ser divididos em partes, por exemplo, um nome pessoal pode conter as partes prenome, sobrenome, datas, nome completo e termo de tratamento;
- suportar componentes individuais em assuntos pré-coordenados, hierarquias geográficas e nomes/títulos: as partes que constituem estes rótulos podem ser identificadas individualmente (com um URI próprio), por exemplo, no ponto de acesso de assunto “Estados Unidos – História – Guerra Civil, 1861-1865” os três componentes terão seus próprios URIs (LIBRARY OF CONGRESS, 2015).

Ainda sobre sua relação com o SKOS,

O MADS/RDF é projetado para complementar o SKOS e, desse modo, está formalmente mapeado para o vocabulário do SKOS em RDF. Esse mapeamento, que pode ser utilizado para o propósito de inferência, facilita o intercâmbio de dados entre um usuário de MADS/RDF e um usuário de SKOS. Porém, o MADS/RDF é um modelo de dados mais especificamente definido para representar as complexidades dos dados de autoridade, de vocabulários e de tesouros. (LIBRARY OF CONGRESS, 2015, tradução nossa).

No MADS/RDF as principais classes são *madsrdf:Authority*, que inclui recursos com rótulos autorizados, e *madsrdf:Variant*, que inclui recursos com rótulos variantes

(não autorizados). Os recursos dessas classes são relacionados pela propriedade *madsrdf:hasVariant*. As relações possibilitadas pelo SKOS são possíveis também com o MADS/RDF. Algumas das classes e propriedades do MADS/RDF estão exemplificadas na Figura 20.

Figura 20 – Exemplos de classes e propriedades do MADS/RDF

Exemplo 1

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .

<http://arquivodeautoridade.com/entidade/12345>
  rdf:type madsrdf:Authority ;
  rdf:type madsrdf:PersonalName ;
  madsrdf:authoritativeLabel "Assis, Machado de, 1839-1908" ;
  madsrdf:hasVariant [
    rdf:type madsrdf:Variant ;
    rdf:type madsrdf:PersonalName ;
    madsrdf:variantLabel "Assis, Joaquim Maria Machado de, 1839-1908" ] ;
  madsrdf:birthDate "1839-06-21" ;
  madsrdf:deathDate "1908-09-29" ;
  madsrdf:gender "Masculino"@pt ;
  madsrdf:associatedLocale <http://arquivodeautoridade.com/local/13579> .
```

Exemplo 2

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .

<http://tesouro.com.br/conceitos/0256>
  rdf:type madsrdf:Authority ;
  rdf:type madsrdf:ComplexSubject ;
  madsrdf:authoritativeLabel "Animais domésticos - Comercialização"@pt ;
  madsrdf:componentList [
    rdf:first <http://tesouro.com.br/conceitos/0031> ;
    rdf:rest [
      rdf:first [
        rdf:type madsrdf:Topic ;
        rdf:type madsrdf:Authority ;
        madsrdf:authoritativeLabel "Comercialização"@pt ] ;
      rdf:rest rdf:nil ]
  ] .

<http://tesouro.com.br/conceitos/0031>
  rdf:type madsrdf:Authority ;
  rdf:type madsrdf:Topic ;
  madsrdf:authoritativeLabel "Animais domésticos"@pt ;
  madsrdf:hasVariant [
    rdf:type madsrdf:Topic ;
    rdf:type madsrdf:Variant ;
    madsrdf:variantLabel "Animais de criação"@pt ] ;
```

madsrdf:hasBroaderAuthority <http://tesouro.com.br/conceitos/0028> ;
 madsrdf:hasNarrowerAuthority <http://tesouro.com.br/conceitos/0035> ;
 madsrdf:hasExactExternalAuthority <http://id.loc.gov/authorities/subjects/sh85038908> .

Fonte: Elaborada pelo autor.

As classes e propriedades exemplificadas na Figura 20 possibilitam, entre outros:

- declarar que um recurso é um ponto de acesso autorizado (*madsrdf:Authority*);
- declarar que um recurso representa um nome pessoal (*madsrdf:PersonalName*) ou um termo tópico (*madsrdf:Topic*);
- declarar que um ponto de acesso resulta da coordenação de outros pontos de acesso (*madsrdf:ComplexSubject*);
- declarar a cadeia de caracteres que compõem um ponto de acesso autorizado (*madsrdf:authoritativeLabel*);
- declarar que ao ponto de acesso autorizado estão relacionados pontos de acesso não autorizados (*madsrdf:hasVariant*);
- declarar valores para propriedades como data de nascimento (*madsrdf:birthDate*), data de morte (*madsrdf:deathDate*), gênero (*madsrdf:gender*) e local associado (*madsrdf:associatedLocale*);
- relacionar um termos a termos mais genéricos (*madsrdf:hasBroaderAuthority*) e a termos mais específicos (*madsrdf:hasNarrowerAuthority*);
- relacionar o recurso a recursos equivalentes em outros conjuntos de dados (*madsrdf:hasExactExternalAuthority*).

O MADS/RDF relaciona-se estreitamente com o Formato MARC 21 para Dados de Autoridade, uma vez que deriva desse formato (LIBRARY OF CONGRESS, 2015). O MADS/XML, que serviu de base para o MADS/RDF, foi criado como um subconjunto dos elementos existentes no Formato MARC 21, de modo que nem todos os dados descritos com este formato podem ser descritos com o MADS/XML (ASSUMPCÃO; SANTOS, 2013b, p. 121) e, conseqüentemente, com o MADS/RDF.

Alguns dos dados do Formato MARC 21 que não podem ser descritos com o MADS/RDF têm por objetivo auxiliar no uso e na gestão dos dados de autoridade, são exemplos desses dados: datas de criação e alteração dos dados, instituição responsável pelos dados, regras utilizadas, etc. Essa limitação do MADS/RDF, no entanto, não se caracteriza necessariamente como uma desvantagem, uma vez que é suprida pelo uso de

classes e/ou de propriedades de outros vocabulários. Nesse sentido, entende-se que a consideração acerca do MADS/XML, apresentada por Assumpção e Santos (2013b, p. 123), também se aplica ao MADS/RDF:

O nível de especificidade provido pelo MADS mostra-se suficiente para a criação de registros de autoridade com o nível de funcionalidade requerido por grande parte das agências catalogadoras, por exemplo, as bibliotecas universitárias. Esse nível de funcionalidade muitas vezes inclui somente o ponto de acesso autorizado, as remissivas “ver” e “ver também” e alguma nota de escopo ou fontes de informação consultadas para a criação do registro.

Assim, considerando seu nível de especificidade e sua similaridade com o Formato MARC 21 para Dados de autoridade, entende-se que o MADS/RDF pode ser de interesse para as instituições que desejam publicar seus dados de autoridade utilizando um vocabulário que atenda às necessidades dos registros de autoridade atualmente utilizados no domínio bibliográfico.

4.1.3 RDA Element Sets⁵²

Em 2010, como resultado de estudos iniciados principalmente a partir de 1997, foi publicado o *Resource Description and Access* (RDA), um substituto do AACR2r. Tendo em sua base os modelos conceituais FRBR, FRAD e FRSAD, o RDA apresenta diretrizes e instruções para o registro de atributos e de relacionamentos de entidades de interesse do domínio bibliográfico (obra, expressão, manifestação, item, pessoa, família, entidade coletiva, conceito, objeto, evento e lugar) (RDA STEERING COMMITTEE, 2016). Segundo Hillmann et al. (2010, tradução nossa),

O RDA essencialmente padroniza como o conteúdo dos metadados é identificado, transcrito e geralmente estruturado, embora ele seja independente de qualquer codificação de metadados específica. O RDA também identifica um conjunto geral de elementos de metadados e, em muitos casos, provê um vocabulário controlado para o uso como conteúdo de um elemento.

Durante o desenvolvimento do RDA, foram criados também vocabulários de classes, de propriedades e de valores a partir das entidades, dos atributos e dos relacionamentos presentes no RDA, o que resultou nos *RDA Element Sets* (Conjuntos de

⁵² O conteúdo desta subseção baseia-se no trabalho “*RDA Element Sets and RDA Value Vocabularies: Vocabularies for Resource Description in the Semantic Web*” (ASSUMPCÃO; SANTAREM SEGUNDO; SANTOS, 2015), publicado nos anais da *9th International Conference on Metadata and Semantics Research* (MTSR).

elementos do RDA) e nos *RDA Value Vocabularies* (Vocabulários de valores do RDA), que começaram a ser publicados em 2014 na plataforma *Open Metadata Registry*⁵³ e, posteriormente, no *RDA Registry*⁵⁴ (ASSUMPÇÃO; SANTAREM SEGUNDO; SANTOS, 2015, p. 150).

Os *RDA Element Sets*, que ainda se encontram em desenvolvimento, atualmente incluem um vocabulário com as classes, seis vocabulários com propriedades restritas a cada uma das classes, um vocabulário com propriedades não restritas às classes, um vocabulário com propriedades administrativas e um vocabulário criado a partir do *RDA/ONIX Framework for Resource Categorization*⁵⁵.

O primeiro vocabulário inclui as classes Obra (*rdac:C10001*), Expressão (*rdac:C10006*), Manifestação (*rdac:C10007*), Item (*rdac:C10003*), Agente (*rdac:C10002*), Pessoa (*rdac:C10004*), Família (*rdac:C10008*), Entidade coletiva (*rdac:C10005*) e Lugar (*rdac:C10009*), definidas no *namespace* com o URI <http://rdaregistry.info/Elements/c/>. Diferentemente do SKOS e do MADS/RDF, em que são empregadas palavras em língua inglesa para compor os URIs das classes e das propriedades, nos *RDA Element Sets* e nos *RDA Value Vocabularies* são utilizados códigos alfanuméricos, o que torna os URIs independentes de qualquer idioma.⁵⁶

Para cada uma das classes Obra, Expressão, Manifestação, Item, Agente e Lugar há um vocabulário de propriedades criado a partir dos atributos e relacionamentos presentes no RDA: a propriedade “idioma da expressão” (*rdac:P20006*), por exemplo, foi criada a partir do atributo “*Language of Expression*” (RDA STEERING COMMITTEE, 2016, 6.11). Em cada vocabulário, as propriedades têm seu domínio (*rdfs:domain*) declarado; as propriedades destinadas à Manifestação, por exemplo, têm como domínio a classe Manifestação. Em alguns casos estão declarados também *ranges*: a propriedade “tem autor”, por exemplo, possui como *range* a classe Agente, o que indica que seu valor é um recurso que pertence a essa classe.

Não há vocabulários específicos para as classes Pessoa, Família e Entidade coletiva, pois, por serem subclasses da classe Agente, elas utilizam as propriedades do vocabulário definido para essa classe. Há, no entanto, propriedades aplicáveis a toda

⁵³ Disponível em: <<http://metadataregistry.org/>>. Acesso em: 17 ago. 2016.

⁵⁴ Disponível em: <<http://www.rdaregistry.info/>>. Acesso em: 17 ago. 2016.

⁵⁵ Disponível em: <<https://www.loc.gov/marc/marbi/2007/5chair10.pdf>>. Acesso em: 17 ago. 2016.

⁵⁶ Às classes e às propriedades são atribuídos também URIs léxicos formados por palavras em língua inglesa, por exemplo, <http://rdaregistry.info/Elements/e/languageOfExpression.en>. Esses URIs, no entanto, apontam para os URIs com códigos alfanuméricos e estão sujeitos a alterações.

classe Agente e propriedades aplicáveis apenas a uma de suas subclasses. A propriedade “é cantor de” (*rdaa:P50084*), por exemplo, tem como domínio a classe Agente, já a propriedade “tem data de nascimento” (*rdaa:P50121*) tem como domínio a classe Pessoa.

Essas restrições de domínio e de *range* fazem com que as propriedades sejam utilizadas apenas nos contextos em que o modelo conceitual FRBR está sendo utilizado em alguma medida. As propriedades desses vocabulários são consideradas, portanto, propriedades restritas ao FRBR. O modelo conceitual FRBR foi desenvolvido no domínio bibliográfico e, mesmo assim, não tem sido amplamente adotado por este no desenvolvimento de seus sistemas de informação. Sobre isso, Marcondes (2012, p. 183) comenta que, embora o modelo conceitual tenha sido incorporado ao RDA, “[...] a adoção do FRBR ainda é lenta, dado o legado de milhões de registros representados em formato MARC e segundo as regras do CCAA2 [AACR2r]”.

Assim, entende-se aqui que a elaboração de vocabulários de propriedades para uso exclusivamente em sistemas baseados no FRBR não favoreceria de forma significativa a utilização do RDA além do domínio bibliográfico ou mesmo dentro dele.

A partir da necessidade de superar a restrição pelo uso do FRBR e de criar vocabulários que pudessem ser utilizados na Web Semântica como um todo, foi criado um vocabulário com propriedades não restritas ao FRBR. As propriedades desse vocabulário também derivam dos atributos e dos relacionamentos do RDA, no entanto, para elas não foram declarados domínios e nem *ranges*.

O vocabulário de propriedades administrativas (referidas como meta-elementos), identificado pelo URI <http://rdaregistry.info/Elements/z/>, inclui propriedades aplicadas na descrição de quaisquer recursos: fonte consultada (*rdaz:P60001*), nota do catalogador (*rdaz:P60002*), status de identificação (*rdaz:P60003*), explicação do relacionamento (*rdaz:P60004*), data de uso (*rdaz:P60005*) e escopo de uso (*rdaz:P60006*), não havendo, portanto, declarações de domínios e *ranges*. O vocabulário *RDA/ONIX Framework element set* inclui classes e propriedades sem domínios e sem *ranges* representando atributos e entidades do *RDA/ONIX Framework for Resource Categorization*.

Das propriedades definidas nos *RDA Element Sets* estão exemplificadas na Figura 21 algumas que se aplicam à classe Agente e as suas subclasses (Pessoa, Família e Entidade coletiva) e algumas do vocabulário de propriedades administrativas.⁵⁷

Figura 21 – Exemplos de propriedades dos *RDA Element Sets*

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdaa: <http://rdaregistry.info/Elements/a/> .
@prefix rdac: <http://rdaregistry.info/Elements/c/> .
@prefix rdaz: <http://rdaregistry.info/Elements/z/> .

<http://pessoas.com/entidade/12345>
  rdf:type rdac:C10004 ;
  rdaa:P50094 "Assis, Machado de, 1839-1908" ;
  rdaa:P50117 "Assis, Machado de" ;
  rdaa:P50103 "Assis, Joaquim Maria Machado de" ;
  rdaa:P50121 "1839-06-21" ;
  rdaa:P50120 "1908-09-29" ;
  rdaa:P50116 "Masculino"@pt ;
  rdaa:P50104 "Escritor"@pt ;
  rdaa:P50100 "Literatura brasileira"@pt ;
  rdaa:P50112 <http://arquivodeautoridade.com/entidade/54321> ;
  rdaz:P60001 "Wikipédia, 20 jul. 2016 (Joaquim Maria Machado de Assis; Rio de Janeiro, 21 de junho de 1839 - Rio de Janeiro, 29 de setembro de 1908; foi um escritor brasileiro)" .

<http://arquivodeautoridade.com/entidade/54321>
  rdf:type rdac:C10005 ;
  rdaa:P50006 "Academia Brasileira de Letras" ;
  rdaa:P50041 "Academia Brasileira de Letras" ;
  rdaa:P50025 "ABL" ;
  rdaa:P50037 "1897" ;
  rdaa:P50029 <http://pessoas.com/entidade/12345> .

```

Fonte: Elaborada pelo autor.

As propriedades exemplificadas na Figura 21 permitem:

- declarar que um recurso é uma pessoa ou uma entidade coletiva;
- declarar o ponto de acesso autorizado para uma pessoa (*rdaa:P50094*) ou uma entidade coletiva (*rdaa:P50006*);
- declarar propriedades como o nome preferido (*rdaa:P50117*), os nomes variantes (*rdaa:P50103*), as datas de nascimento (*rdaa:P50121*) e de morte (*rdaa:P50120*), o gênero (*rdaa:P50116*), a profissão (*rdaa:P50104*) e o campo de atividade (*rdaa:P50100*) de uma pessoa;

⁵⁷ Exemplos de outras propriedades são dados por Assumpção, Santarem Segundo e Santos (2015).

- declarar propriedades como o nome preferido (*rdaa:P50041*), os nomes variantes (*rdaa:P50025*) e a data de fundação (*rdaa:P50037*) de uma entidade coletiva;
- relacionar as entidades, por exemplo, declarar que uma pessoa fundou uma entidade coletiva (*rdaa:P50112*) e que uma entidade coletiva foi fundada por uma pessoa (*rdaa:P50029*);
- declarar quais fontes foram consultadas para a criação dos dados de autoridade (*rdaz:P60001*).

Além dos vocabulários de classes e de propriedades, foram criados a partir do RDA vocabulários de valores. Alguns atributos do RDA têm seus valores definidos por listas de termos, exemplos desses atributos são “*Carrier type*”, “*Content type*” e “*Illustrative content*” (RDA STEERING COMMITTEE, 2016, 3.3, 6.9, 7.15). Embora os vocabulários de valores tenham sido criados a partir de atributos do RDA e tais atributos tenham sido utilizados também para a criação de propriedades, não há a obrigatoriedade do uso de um vocabulário de valores com sua respectiva propriedade ou vice-versa. Por exemplo, não é obrigatório que os valores do vocabulário *Content Type* sejam utilizados com a propriedade “*has content type*”, da mesma forma que não é obrigatório que essa propriedade tenha como valores apenas os termos do vocabulário *Content Type*.

O RDA tem se apresentado mais apto para a criação de registros de autoridade representando pessoas, famílias e entidades coletivas, se comparado com seu antecessor, o AACR2r (ASSUMPÇÃO; SANTOS, 2013a, p. 222), o que, nesta pesquisa, considera-se uma vantagem do novo código. Nesse sentido, destaca-se que a proximidade dos vocabulários com o próprio RDA pode favorecer sua adoção por parte das instituições que utilizam esse código de catalogação. Além disso, os mapeamentos já existentes entre os atributos e relacionamentos do RDA e o Formato MARC 21 para Dados de Autoridade auxiliam no mapeamento dos campos e subcampos desse formato para as propriedades dos vocabulários.

Outro ponto de destaque é que, sem a necessidade de utilização das entidades definidas no FRBR, o vocabulário de propriedades não restritas tem suas possibilidades de uso ampliadas, podendo ser utilizado tanto no domínio bibliográfico quanto fora dele.⁵⁸

⁵⁸ Outras considerações acerca das vantagens e das desvantagens dos *RDA Element Set* e dos *RDA Value Vocabularies* são apresentadas por Assumpção, Santarem Segundo e Santos (2015, p. 157).

A apresentação dos vocabulários realizada nesta seção não teve por objetivo comparar tais vocabulários ou indicar o mais apropriado para a publicação de dados de autoridade como *Linked Data*, mas sim de ilustrar algumas das possibilidades atuais. Sobre a variedade de vocabulários para a descrição de um mesmo tipo de recurso, Santos e Alves (2009) destacam que “A tendência é a coexistência de vários formatos de metadados, pois cada um apresenta uma característica específica que atende à necessidade de descrição de determinados tipos de recursos informacionais e comunidades de interesses distintos.” Ao passo que Serra e Santarem Segundo (2017, p. 183) entendem que

Um ponto que pode ser desafiador aos bibliotecários será a identificação dos instrumentos que serão adotados para favorecer a interligação de dados. Devem ser selecionadas as ontologias adequadas, assim como os catálogos de autoridades, cuja seleção deve ser orientada pelo escopo da instituição e de seu público usuário.

No domínio bibliográfico, em especial, nos processos descritivos da catalogação, sempre se buscou a utilização de padrões de metadados universais, de modo a facilitar o intercâmbio e o reuso de dados. No âmbito da Web Semântica, a utilização de um mesmo padrão de metadados, ou melhor, de um mesmo vocabulário de classes e de propriedades, deixa de ser o principal requisito para o intercâmbio e o reuso dos dados. Ocupando essa posição encontra-se o modelo de dados RDF que, por ter como unidade básica uma tripla, possibilita a descrição de recursos utilizando, simultaneamente, classes e propriedades de diferentes vocabulários, assim como possibilita, a partir de recursos do RDFS e da OWL, um “caminhar” por entre essas classes e propriedades para a realização de inferências acerca dos dados. Assim, a variedade de vocabulários para a descrição de um mesmo tipo de recurso, no caso, dados de autoridade, não representa uma falta de consenso da área ou um entrave à publicação de *Linked Data*, mas sim uma vantagem, uma vez que classes e propriedades de diferentes vocabulários podem ser reunidas de modo a atender as necessidades de descrição específicas de determinada instituição. A possibilidade de combinação de diferentes vocabulários é ilustrada na seção seguinte, em que são apresentados conjuntos de dados de autoridade publicados como *Linked Data*.

4.2 Dados de autoridade publicados como *Linked Data*

Ao representarem os conjuntos de dados publicados como *Linked Open Data* em uma *Linking Open Data cloud* (Figura 10), Abele et al. (2017) agrupam os conjuntos de dados do domínio bibliográfico na categoria “Publicações”. Nessa categoria são identificados conjuntos de dados de autoridade.

A proposição de um modelo para a publicação de dados de autoridade como *Linked Data* a partir da conceituação das funções dos dados de autoridade, objetivo geral desta pesquisa, requer, entre outros objetivos específicos, a identificação das principais iniciativas para a publicação de dados de autoridade como *Linked Data*, acompanhada da descrição das políticas, dos procedimentos e das tecnologias nelas utilizadas. Desse modo, nesta seção são apresentadas algumas iniciativas para a publicação de dados de autoridade como *Linked Data* que figuram entre os conjuntos de dados elencados por Abele et al. (2017), são elas: o *Library of Congress Linked Data Service* (Seção 4.2.1), o portal *datos.bne.es* (Seção 4.2.2), o projeto *data.bnf.fr* (Seção 4.2.3), o *Virtual International Authority File* (VIAF) (Seção 4.2.4) e o *AGROVOC Multilingual agricultural thesaurus* (Seção 4.2.5).

4.2.1 *Library of Congress Linked Data Service*

O *Library of Congress Linked Data Service*⁵⁹ tem em seu escopo “[...] prover acesso aos padrões comumente encontrados e a vocabulários aprovados pela *Library of Congress*” (LIBRARY OF CONGRESS, 2016b, tradução nossa). O serviço, iniciado em 2009, cobria primeiramente a LCSH (GUENTHER, 2011). Nos anos seguintes, foram acrescentados outros conjuntos de dados: o arquivo de autoridade de nomes de pessoas e entidades coletivas e de títulos de obras (*LC Name Authority File*); tesouros, listas de cabeçalhos de assunto e listas de termos representando gêneros, formas, meios de performance e grupos de pessoas (*LC Children's Subject Headings, LC Medium of Performance Thesaurus for Music, Thesaurus for Graphic Materials, AFS Ethnographic Thesaurus, LC Genre/Form Terms, LC Demographic Group Terms*), o sistema de classificação da LC (a *LC Classification*), listas de códigos de instituições, idiomas,

⁵⁹ Disponível em: <<http://id.loc.gov/>>. Acesso em: 12 dez. 2017.

designadores de relacionamento, países e áreas geográficas (*Cultural Heritage Organizations, MARC Relators, MARC Countries, MARC Geographic Areas, MARC Languages, ISO639-1 Languages, ISO639-2 Languages, ISO639-5 Languages*), além de esquemas e listas de códigos menores e vocabulários voltados à preservação (derivados do padrão *Preservation Metadata Maintenance Activity (PREMIS)*) (LIBRARY OF CONGRESS, 2016a).

Para cada conjunto de dados, foi definido um padrão para a construção do URI. Para o arquivo de autoridade de nomes de pessoas e entidades coletivas e de títulos de obras é utilizado o URI base *http://id.loc.gov/authorities/names/*, ao qual é acrescentado o número de controle atribuído pela LC ao registro de autoridade, por exemplo, o número de controle do registro de autoridade representando a pessoa Machado de Assis é *n80002329*, portanto seu URI é *http://id.loc.gov/authorities/names/n80002329*. Os URIs dos conceitos da LCSH são construídos de forma semelhante tendo como base o URI *http://id.loc.gov/authorities/subjects/*. Alguns exemplos de URIs dos diferentes conjuntos de dados são:

- *http://id.loc.gov/authorities/names/n80002329*, que identifica a pessoa Machado de Assis no *LC Name Authority File*;
- *http://id.loc.gov/authorities/subjects/sh85038908*, que identifica o conceito de animais domésticos na LCSH;
- *http://id.loc.gov/vocabulary/relators/trl*, que identifica a função de tradutor na *MARC List of Relator Terms*; e
- *http://id.loc.gov/vocabulary/languages/por*, que identifica o idioma português na *MARC List for Languages*.

Os dados publicados pelo serviço estão disponíveis em domínio público (LIBRARY OF CONGRESS, 2016b) e representados utilizando o SKOS, o MADS/RDF, a OWL e a *Ontology for Library of Congress Classification*,⁶⁰ criada para atender as necessidades de descrição específicas do sistema de classificação da LC. No início do serviço, que contava somente com os dados da LCSH, era utilizado apenas o SKOS. No entanto, a necessidade de maior especificidade na representação dos dados levou a LC ao desenvolvimento e ao uso do MADS/RDF:

⁶⁰ Disponível em: <<http://id.loc.gov/ontologies/lcc>>. Acesso em: 12 dez. 2017.

A *Library of Congress* tem experiência na representação de dados de autoridade em SKOS/RDF desde o início de 2009, antes da especificação do SKOS tornar-se uma recomendação formal do W3C. A LC foi um dos primeiros a utilizar o SKOS. O SKOS foi empregado em seu serviço de *Linked Data*, ID.LOC.GOV, que disponibilizou toda a *LC Subject Headings* (LCSH) no formato SKOS/RDF. Essa experiência enriquecedora mostrou, imediatamente após seu lançamento e no decorrer do tempo, alguns dos desafios envolvidos na representação de dados de autoridade mais complexos utilizando somente o SKOS. Uma das grandes questões rapidamente trazidas à atenção da LC pelos usuários dos dados foi: a LC não está transmitindo dados úteis em seus registros (MARC) de autoridade, pois o SKOS não provê um modo de capturar as informações, tais como os múltiplos componentes, e seus tipos, de cabeçalhos de assunto pré-coordenados. (LIBRARY OF CONGRESS, 2015, tradução nossa).

Os dados estão disponíveis pela interface web do serviço (páginas HTML) e pela negociação de conteúdo, por meio das quais os dados são obtidos serializados em XML, *N-Triples* e JSON. Os conjuntos de dados estão disponíveis também para *download* serializados em XML, *Turtle* e *N-Triples*. Atualmente não há um SPARQL *endpoint* para a consulta aos dados (LIBRARY OF CONGRESS, 2016c).

Não foi localizada uma relação dos conjuntos de dados aos quais os dados do *Library of Congress Linked Data Service* estão relacionados. No entanto, durante a utilização do serviço é possível encontrar *links* para recursos do VIAF, da *Deutschen Nationalbibliothek* (DNB)⁶¹, do *data.bnf.fr*, do *Nuovo soggettario*⁶² e do *National Agricultural Library Thesaurus*⁶³.

4.2.2 *datos.bne.es*

O portal *datos.bne.es*⁶⁴ é um projeto da *Biblioteca Nacional de España* (BNE) e do *Ontology Engineering Group*, da *Universidad Politécnica de Madrid*. Iniciado em 2011, o projeto tem por objetivo servir tanto aos usuários finais da BNE quanto aos desenvolvedores que trabalham com as tecnologias da Web Semântica (BIBLIOTECA NACIONAL DE ESPAÑA, 2016a, tradução nossa):

⁶¹ Disponível em: <http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkedata_node.html>. Acesso em: 23 ago. 2016.

⁶² Disponível em: <<http://thes.bncf.firenze.sbn.it/>>. Acesso em: 23 ago. 2016.

⁶³ Disponível em: <<https://agclass.nal.usda.gov/>>. Acesso em: 23 ago. 2016.

⁶⁴ Disponível em: <<http://datos.bne.es>>. Acesso em: 23 ago. 2016.

Para o usuário final e para o pesquisador trata-se de um projeto experimental que busca propor uma aproximação e uma exploração de dados bibliográficos totalmente diferentes dos catálogos tradicionais, propondo uma experiência de navegação completamente nova a partir de um enfoque oposto ao tradicional, integrando os distintos recursos da biblioteca e enriquecendo os próprios dados com outros dados externos.

Para o especialista da computação, trata-se do acesso a um banco de dados descrito com linguagem semântica oferecido aos especialistas e desenvolvedores para sua reutilização. Para ele, os dados bibliográficos, de autores e de assuntos foram transformados da arquitetura tradicional das bibliotecas para modelos e estruturas da Web Semântica, tecnologia cujo propósito é expor os dados na web de um modo mais reutilizável e interoperável com outras aplicações, explorando as relações existentes entre os dados, tanto internas quanto externas.

Os dados incluídos no portal foram extraídos de registros bibliográficos e de autoridade da BNE, de modo que contemplam a descrição de pessoas, entidades coletivas, obras, conceitos (representados por cabeçalhos de assunto) e manifestações (BIBLIOTECA NACIONAL DE ESPAÑA, 2015, p. 2; VILA SUERO; VILLAZÓN TERRAZAS; GÓMEZ PÉREZ, 2012, p. 2). A esses dados, foram integrados também dados de fontes externas, tais como imagens dos autores, obtidas da *Wikipedia*, *links* para recursos equivalentes em outros conjuntos de dados e *links* para documentos digitalizados disponíveis na *Biblioteca Digital Hispánica*⁶⁵ (BIBLIOTECA NACIONAL DE ESPAÑA, 2015, p. 3).

A transformação, os *links* e a publicação dos dados ocorreram em sete etapas:

- **etapa de especificação**, em que foram: (1) identificadas e analisadas as fontes de dados (registros bibliográficos e de autoridade da BNE nos Formatos MARC 21), (2) definido um padrão para os URIs (o número de controle do registro bibliográfico ou de autoridade seria acrescentado ao final do URI base *http://datos.bne.es/resource/*) e (3) escolhida a licença *Creative Commons Public Domain Dedication*⁶⁶ para a publicação dos dados;
- **etapa de modelagem**, em que foram analisados e selecionados vocabulários do domínio bibliográfico (os vocabulários FRBR, FRAD, FRSAD e ISBD, publicados pela IFLA, e, adicionalmente, elementos de outros vocabulários como o *RDA Elements Sets*, o DC, o SKOS e o MADS/RDF), a partir dos quais foram selecionadas as classes e as propriedades para a modelagem das entidades e de

⁶⁵ Disponível em: <<http://bdh.bne.es/>>. Acesso em: 23 ago. 2016.

⁶⁶ Disponível em: <<https://creativecommons.org/publicdomain/zero/1.0/>>. Acesso em: 23 ago. 2016.

seus atributos e selecionados os relacionamentos entre as entidades; a modelagem realizada é apresentada de forma genérica na Figura 22;

- **etapa de geração**, que incluiu o mapeamento dos registros bibliográficos e de autoridade para as classes e propriedades dos vocabulários e a transformação desses registros em declarações RDF utilizando o software MARiMba⁶⁷;
- **etapa de relacionamento**, que incluiu a identificação de conjuntos de dados que pudessem ser relacionados aos dados da BNE (*VIAF*, *Library of Congress Linked Data Service*, *DBpedia*, *Gemeinsame Normdatei* da DNB⁶⁸, *Libris*⁶⁹, e *Système Universitaire de Documentation* (SUDOC)⁷⁰), a descoberta e a validação de *links*;
- **etapa de curadoria dos dados**, realizada paralelamente às etapas de especificação, modelagem e geração com o objetivo de assegurar a qualidade dos dados;
- **etapa de publicação**, que incluiu a disponibilização dos dados na Web a partir da publicação do conjunto de dados (provendo uma interface web com negociação de conteúdo, um SPARQL *endpoint* e uma *Application Programming Interface* (API)), de metadados descrevendo o conjunto (utilizando o vocabulário *Vocabulary of Interlinked Datasets* (VOID)⁷¹) e do registro do conjunto de dados no *Datahub*; e
- **etapa de exploração**, que incluiu o desenvolvimento de aplicações e de serviços para a exploração dos dados e o oferecimento de interfaces ricas para os usuários e os desenvolvedores (VILA SUERO; VILLAZÓN TERRAZAS; GÓMEZ PÉREZ, 2012; VILA SUERO; GÓMEZ PÉREZ, 2013; BIBLIOTECA NACIONAL DE ESPAÑA, 2016a, 2016b).

⁶⁷ Disponível em: <<http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/228-marimba/>>. Acesso em: 23 ago. 2016.

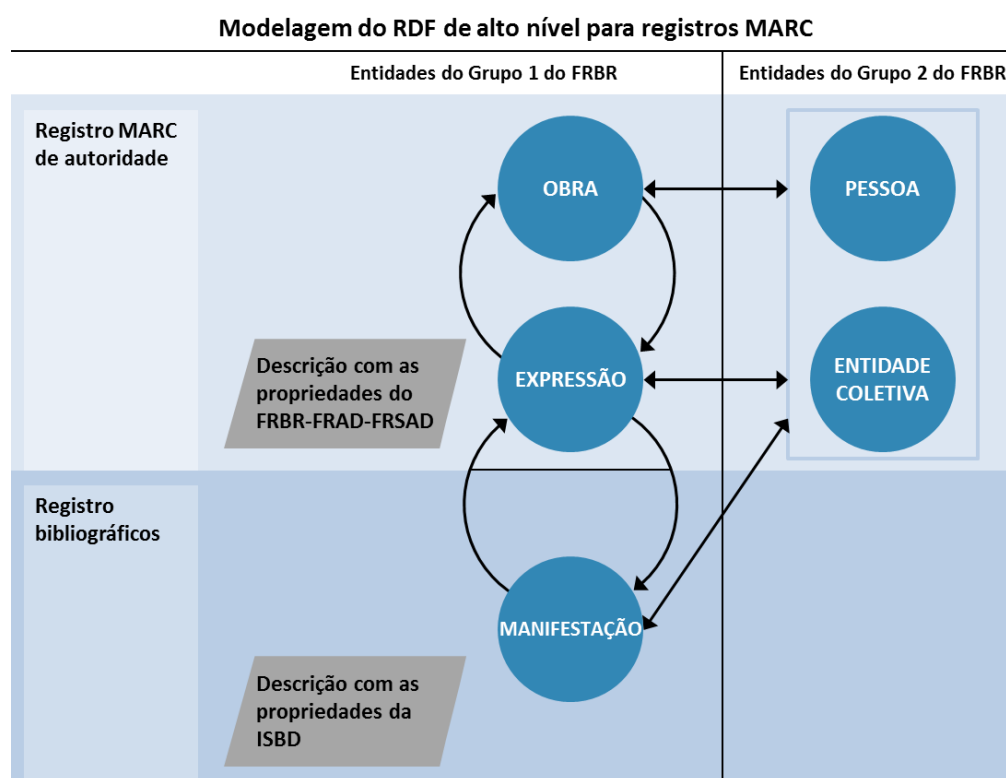
⁶⁸ Disponível em: <<http://www.dnb.de/gnd>>. Acesso em: 23 ago. 2016.

⁶⁹ Disponível em: <<http://libris.kb.se/>>. Acesso em: 23 ago. 2016.

⁷⁰ Disponível em: <<http://www.sudoc.abes.fr/>>. Acesso em: 23 ago. 2016.

⁷¹ Disponível em: <<https://www.w3.org/TR/void/>>. Acesso em: 23 ago. 2016.

Figura 22 – Modelagem utilizada no portal *datos.bne.es*.



Fonte: *Biblioteca Nacional de España* (2016b, tradução nossa).

São exemplos de URIs que identificam pessoas, entidades coletivas, obras e conceitos no portal da BNE:

- <http://datos.bne.es/resource/XX1029037>, que identifica a pessoa Machado de Assis;
- <http://datos.bne.es/resource/XX218146>, que identifica a entidade coletiva Academia Brasileira de Letras;
- <http://datos.bne.es/resource/XX2067619>, que identifica a obra “Dom Casmurro”;
- <http://datos.bne.es/resource/XX525110>, que identifica o conceito de animais domésticos, e <http://datos.bne.es/tema/XX525110.html>, que identifica o registro de autoridade para esse conceito.

Em razão dos procedimentos utilizados, os dados do portal não estão sincronizados com o catálogo da BNE, ou seja, os procedimentos de geração, de relacionamento e de publicação não ocorrem em tempo real, mas sim a cada intervalo de tempo determinado (BIBLIOTECA NACIONAL DE ESPAÑA, 2015, p. 2).

4.2.3 *data.bnf.fr*

O projeto *data.bnf.fr*⁷², desenvolvido pela *Bibliothèque nationale de France* (BnF), publica dados sobre autores, obras e conceitos automaticamente mesclados a partir de diferentes bases de dados da BnF (SIMON et al., 2013, p. 1). Disponível online desde 2011, o projeto objetiva:

- tornar os dados produzidos pela BnF mais visíveis na Web;
- federar os dados produzidos pela BnF, tanto dentro quanto fora dos catálogos;
- contribuir com a colaboração e a troca de metadados por meio da criação de *links* entre recursos estruturados e confiáveis;
- facilitar o reuso de metadados (sob um licença aberta) por terceiros. (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016a, tradução nossa).

As principais fontes de dados da BnF vinculadas ao projeto são: o catálogo principal de publicações (que inclui o arquivo de autoridade de nomes de pessoas, de entidades coletivas e títulos de obras, e o *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU)⁷³), o catálogo do arquivo e da coleção de manuscritos e a biblioteca digital *Gallica*⁷⁴.

Para a disponibilização dos dados foi adotada a *French Open Licence*: uma licença baseada na *Creative Commons Attribution License* (CC-BY), adaptada de acordo com a legislação francesa de copyright, utilizada pelo governo francês na disponibilização de dados governamentais e que permite a livre (*freely*) cópia e o reuso dos dados em RDF com fins lucrativos ou não, sendo obrigatória a citação da fonte (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016a).

A modelagem dos dados, realizada a partir de uma adaptação do modelo conceitual FRBR, resultou em um projeto que apresenta três principais tipos de páginas, relacionados entre si:

- páginas de obras, que incluem dados sobre obras e expressões e *links* para as manifestações (registros bibliográficos dos catálogos e da biblioteca digital);

⁷² Disponível em: <<http://data.bnf.fr/>>. Acesso em: 09 ago. 2016.

⁷³ Disponível em: <<http://rameau.bnf.fr/>>. Acesso em: 09 ago. 2016.

⁷⁴ Disponível em: <<http://gallica.bnf.fr/>>. Acesso em: 09 ago. 2016.

- páginas de autores, que incluem dados de pessoas e entidades coletivas associadas às obras, expressões e manifestações; e
- páginas de conceitos, que incluem dados obtidos a partir do RAMEAU, (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016b; SIMON et al., 2013, p. 3-4)

Os URIs para os recursos descritos no projeto foram baseados nos identificadores persistentes *Archival Resource Key (ARK)*⁷⁵, previamente utilizados para os registros dos catálogos, recursos arquivísticos e objetos digitais da BnF (SIMON et al., 2013, p. 2-3). São exemplos de URIs utilizados no projeto:

- <http://data.bnf.fr/ark:/12148/cb118893083>, que identifica a pessoa Machado de Assis;
- <http://data.bnf.fr/ark:/12148/cb11870561q>, que identifica a entidade coletiva Academia Brasileira de Letras;
- <http://data.bnf.fr/ark:/12148/cb11938233x>, que identifica a obra “Dom Casmurro”;
- <http://data.bnf.fr/ark:/12148/cb11930905p>, que identifica o conceito de animais domésticos.

A representação dos recursos foi realizada com os vocabulários SKOS, DC, FOAF, *RDA Element Sets* e *BnF-onto*, esse último criado para atender as demandas específicas do projeto (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016b; SIMON et al., 2013, p. 12).

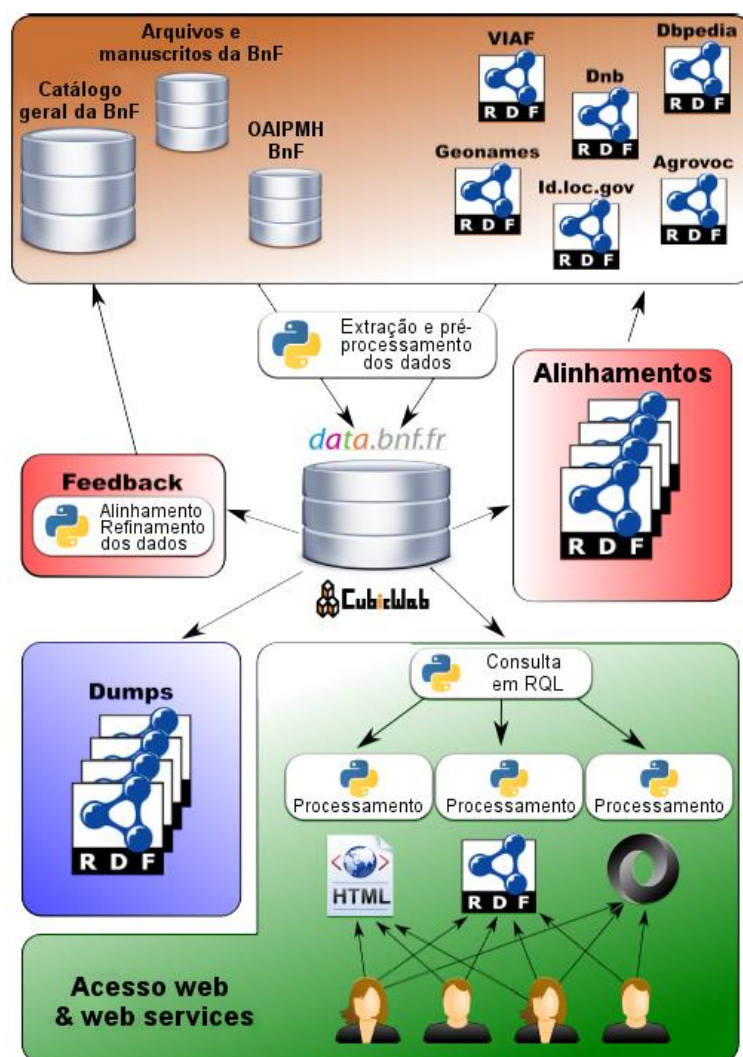
Para extrair os dados das diferentes fontes e formatos, reuni-los e, a partir deles, gerar páginas em HTML e arquivos RDF, foi utilizado o software *CubicWeb*⁷⁶ (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016a). Os dados do projeto foram vinculados aos conjuntos de dados da LC, da DNB, do VIAF, do IdRef,⁷⁷ do *Geonames*, da *DBpedia*, do AGROVOC e do tesouro do *Archives de France* (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2016a), como é observado na Figura 23, que ilustra a arquitetura global do projeto.

⁷⁵ Disponível em: <<https://confluence.ucop.edu/display/Curation/ARK>>. Acesso em: 09 ago. 2016.

⁷⁶ Disponível em: <<https://www.cubicweb.org/>>. Acesso em: 09 ago. 2016.

⁷⁷ Disponível em: <<https://www.idref.fr/>>. Acesso em: 09 ago. 2016.

Figura 23 – Arquitetura global do projeto *data.bnf.fr*



Fonte: Simon et al. (2013, p. 5, tradução nossa).

Os dados estão disponíveis em páginas HTML e em arquivos RDF (serializados nos formatos XML, N3, *N-Triples*, JSON-LD e JSON), obtidos um-a-um, a partir das páginas HTML ou da negociação de conteúdo, ou em conjunto, a partir do *download* de arquivos separados contendo os dados sobre as obras, os autores e os conceitos.

4.2.4 *Virtual International Authority File (VIAF)*

O *Virtual International Authority File (VIAF)*⁷⁸ é um serviço mantido pela *Online Computer Library Center (OCLC)* que combina dados de arquivos de autoridade amplamente utilizados e disponibiliza tais dados na Web com o objetivo de reduzir o custo e aumentar a utilidade dos arquivos de autoridade das bibliotecas. O VIAF teve início em 2003 como uma parceria entre a LC, a DNB e a OCLC e conta atualmente com a participação de mais de cinquenta conjuntos de dados, entre eles, dados das bibliotecas nacionais da Alemanha, Austrália, Brasil, Canadá, Chile, Coréia do Sul, Croácia, Espanha, Estados Unidos, França, Hungria, Irlanda, Israel, Letônia, Líbano, Luxemburgo, Marrocos, Noruega, Países Baixos, Polônia, Portugal, República Tcheca, Rússia, Suécia, Suíça e Taiwan, e de projetos como o *Wikidata*⁷⁹ e o *International Standard Name Identifier (ISNI)*⁸⁰ (ONLINE COMPUTER LIBRARY CENTER, 2016a, 2016b).

Atualmente, o VIAF inclui dados de autoridade sobre pessoas, entidades coletivas, áreas geográficas, obras e expressões. Esses dados são obtidos a partir da reunião dos arquivos de autoridade das instituições que participam do serviço e da junção de todos os registros que representam uma mesma entidade em um “super-registro” de autoridade (ONLINE COMPUTER LIBRARY CENTER, 2014). Um super-registro de autoridade, diferentemente de um registro de autoridade “tradicional” pode conter mais de um ponto de acesso autorizado para a entidade por ele representada, o que ocorre, por exemplo, quando são reunidos registros de autoridade de diferentes idiomas ou alfabetos ou criados a partir de diferentes regras de catalogação.

Para cada super-registro de autoridade é criado um URI a partir do identificador numérico que o identifica dentro do VIAF. Os identificadores dos registros de autoridade de cada instituição que foram reunidos para compor o super-registro também são preservados. Com isso, *links* são estabelecidos entre o VIAF e os arquivos de autoridade das instituições. Nos casos das instituições que possibilitam o acesso aos seus registros por meio de URIs HTTP, por exemplo, a BNE, a BnF, a DNB, a LC e o *Wikidata*, esses links são apresentados nas páginas HTML do VIAF. São exemplos de URIs do VIAF que identificam pessoas, entidades coletivas, áreas geográficas, obras e expressões:

⁷⁸ Disponível em: <<http://viaf.org/>>. Acesso em: 09 ago. 2016.

⁷⁹ Disponível em: <<https://www.wikidata.org/>>. Acesso em: 09 ago. 2016.

⁸⁰ Disponível em: <<http://www.isni.org/>>. Acesso em: 09 ago. 2016.

- <http://viaf.org/viaf/95151633>, que identifica a pessoa Machado de Assis;
- <http://viaf.org/viaf/139117438>, que identifica a entidade coletiva Academia Brasileira de Letras;
- <http://viaf.org/viaf/154512699>, que identifica a cidade do Rio de Janeiro;
- <http://viaf.org/viaf/206860700>, que identifica a obra “Dom Casmurro”;
- <http://viaf.org/viaf/267145542817896642288>, que identifica a expressão em espanhol, realizada por Pablo Cardellino Soto, da obra “Dom Casmurro”;

Para a representação dos super-registros de autoridade em RDF são utilizados os vocabulários SKOS, FOAF, *Schema*⁸¹, *BiblioGraph* (um extensão do Schema)⁸² e DC. Para a descrição do conjunto de dados do VIAF é utilizado o VoID. Os dados, que estão disponíveis sob a *Open Data Commons Attribution License* (ODC-BY), são obtidos em RDF pela negociação de conteúdo, pelas páginas HTML dos super-registros de autoridade (por meio do *download* de arquivos contendo a serialização do RDF em XML) ou por meio do *download* de todo o conjunto de dados do VIAF. Não há um SPARQL *endpoint* para a consulta aos dados.

4.2.5 AGROVOC Multilingual agricultural thesaurus

O *AGROVOC Multilingual agricultural thesaurus*, ou simplesmente AGROVOC, “[...] é um vocabulário controlado cobrindo todas as áreas de interesse da Organização das Nações Unidas para Alimentação e Agricultura (*Food and Agriculture Organization – FAO*, em inglês), incluindo alimentos, nutrição, agricultura, pesca, silvicultura, meio ambiente, etc.” (AGRICULTURAL INFORMATION MANAGEMENT STANDARDS, 2016a, tradução nossa). O AGROVOC foi publicado pela primeira vez, em formato impresso, no início da década de 1980; em 2000 passou a ser mantido apenas em sua versão digital; em 2004 foi convertido para OWL e, em 2009, para o SKOS (CARACCILO et al., 2013, p. 341; AGRICULTURAL INFORMATION MANAGEMENT STANDARDS, 2016b).

Atualmente incluindo mais de 32 mil conceitos com rótulos em até 21 idiomas (AGRICULTURAL INFORMATION MANAGEMENT STANDARDS, 2016b), o AGROVOC, por ter em sua base o SKOS, está modelado a partir:

⁸¹ Disponível em: <<http://schema.org/>>. Acesso em: 09 ago. 2016.

⁸² Disponível em: <<http://bibliograph.net/>>. Acesso em: 09 ago. 2016.

- dos conceitos que, identificados por URIs, são os recursos representados no domínio do AGROVOC; o URI http://aims.fao.org/aos/agrovoc/c_2356, por exemplo, identifica o conceito de animais domésticos;
- dos termos (rótulos) que são utilizados para rotular os conceitos, podendo ser termos preferidos ou não preferidos; por exemplo, “Animal doméstico”, “*Domestic animals*”, “*Animales domésticos*”, “*Animal domestique*” estão declarados como termos preferidos para o conceito de animais domésticos em português, inglês, espanhol e francês, respectivamente; e
- das relações hierárquicas (entre conceitos genéricos e conceitos específicos), não hierárquicas (entre conceitos relacionados) e entre termos; por exemplo, “conceito genérico”, “conceito específico”, “conceito relacionado”, “inclui”, “é usado como”, “causa”, “tem componente”, etc.

Os vocabulários utilizados no AGROVOC são o SKOS (para a declaração das relações entre conceitos, das notações, etc.), sua extensão SKOS-XL (para a declaração de relações entre os termos) e o *Agrontology*,⁸³ criado para atender as necessidades específicas do domínio do AGROVOC (para a declaração de relações entre conceitos não cobertas pelo SKOS). Para a descrição do conjunto de dados do AGROVOC é utilizado o VoID (AGRICULTURAL INFORMATION MANAGEMENT STANDARDS, 2016b; CARACCILO et al., 2013, p. 342-343).

Por meio das propriedades do SKOS destinadas ao relacionamento entre conceitos de diferentes esquemas (*skos:exactMatch*, *skos:closeMatch*, *skos:narrowMatch* e *skos:broadMatch*), o AGROVOC está relacionado a dezesseis vocabulários também publicados como *Linked Data*, alguns deles de disciplinas específicas, por exemplo, o *National Agricultural Library's Agricultural Thesaurus*, da agricultura, e o *General Multilingual Environmental Thesaurus* (GEMET), de ciências ambientais, e outros de caráter geral, por exemplo, a *DBpedia*, a *LCSH* e o *RAMEAU* (AGRICULTURAL INFORMATION MANAGEMENT STANDARDS, 2015).

Os dados do AGROVOC estão publicados sob a licença *Creative Commons Attribution 3.0 Unported License* (CC-BY 3.0)⁸⁴ e acessíveis pela interface web de busca e de navegação, pelo *download* do conjunto de dados, pelo *web service* e pelo *SPARQL endpoint*. A edição dos dados do AGROVOC e sua transformação em RDF ocorre com o

⁸³ Disponível em: <<http://aims.fao.org/aos/agrontology>>. Acesso em: 09 ago. 2016.

⁸⁴ Disponível em: <<http://creativecommons.org/licenses/by/3.0/>>. Acesso em: 09 ago. 2016.

software *VocBench*⁸⁵, desenvolvido pela FAO. A interface web de busca e de navegação⁸⁶ faz uso do software *Skosmos*⁸⁷.

Nesta seção foram abordadas algumas iniciativas para a publicação de dados de autoridade como *Linked Data*, sendo elas mantidas por bibliotecas nacionais (LC, BNE e BnF) ou por instituições de atuação internacional (OCLC e FAO). Como síntese dessa abordagem, é apresentado o Quadro 4, que contempla as principais características dessas iniciativas.

⁸⁵ Disponível em: <<http://aims.fao.org/vest-registry/tools/vocbench>>. Acesso em: 09 ago. 2016.

⁸⁶ Disponível em: <<http://oek1.fao.org/skosmos/agrovoc/en/>>. Acesso em: 09 ago. 2016.

⁸⁷ Disponível em: <<http://skosmos.org/>>. Acesso em: 09 ago. 2016.

Quadro 4 – Iniciativas para a publicação de dados de autoridade como *Linked Data*

Iniciativa	<i>LC Linked Data Service</i>	<i>datos.bne.es</i>	<i>data.bnf.fr</i>	VIAF	AGROVOC
Abrangência dos dados	Dados sobre pessoas, entidades coletivas, obras, expressões, conceitos (descritores, cabeçalhos de assunto, gêneros, formas, meios de performance, notações de classificação), idiomas, áreas geográficas	Dados sobre pessoas, entidades coletivas, obras, manifestações, e conceitos (cabeçalhos de assunto)	Dados sobre pessoas, entidades coletivas, obras, expressões, manifestações, conceitos	Dados sobre pessoas, entidades coletivas, áreas geográficas, obras e expressões	Dados sobre conceitos (descritores de um tesouro)
Licença	Domínio público	<i>Creative Commons Public Domain Dedication (CC0)</i>	<i>French Open Licence</i>	<i>Open Data Commons Attribution License (ODC-BY)</i>	<i>Creative Commons Attribution 3.0 Unported License</i>
Vocabulários	SKOS, MADS/RDF, OWL, <i>Ontology for Library of Congress Classification</i>	FRBR, FRAD, FRSAD, ISBD, <i>RDA Elements Sets</i> , DC, SKOS, MADS/RDF	SKOS, DC, FOAF, <i>RDA Element Sets</i> , <i>BnF-onto</i>	SKOS, FOAF, <i>Schema</i> , <i>BiblioGraph</i> , DC	SKOS, SKOS-XL, <i>Agrontology</i>
Formas de acesso aos dados	HTML, XML, <i>N-Triples</i> , <i>Turtle</i> , JSON	HTML, XML, <i>Turtle</i> , SPARQL <i>endpoint</i> , API	HTML, XML, N3, <i>N-Triples</i> , JSON-LD, JSON	HTML, XML	HTML, XML, <i>Turtle</i> , JSON-LD, SPARQL <i>endpoint</i> , <i>web service</i>
Conjuntos de dados relacionados	VIAF, DNB, <i>data.bnf.fr</i> , <i>Nuovo soggettario</i> , <i>National Agricultural Library Thesaurus</i>	VIAF, <i>LC Linked Data Service</i> , <i>DBpedia</i> , <i>Gemeinsame Normdatei</i> , <i>Libris</i> , SUDOC	VIAF, <i>LC Linked Data Service</i> , DNB, <i>IdRef</i> , <i>Geonames</i> , <i>DBpedia</i> , AGROVOC, <i>Tesouro do Archives de France</i>	Instituições participantes que atribuem URIs HTTP aos seus registros (BNE, BnF, DNB, LC, <i>Wikidata</i> , etc.)	16 vocabulários entre específicos e gerais (por exemplo, <i>DBpedia</i> , LCSH, RAMEAU)

Fonte: Elaborado pelo autor.

As iniciativas aqui relatadas, independentemente de seu escopo, buscam a inserção de dados de autoridade em um cenário global, inserção essa que passa, obrigatoriamente, por etapas como a definição de identificadores (os URIs), a representação dos dados utilizando um modelo comum (o RDF), o estabelecimento de *links* para outros conjuntos de dados e a publicação dos dados.

Embora aparentem despontar de ideias recentes, as iniciativas no domínio bibliográfico para a criação de bases de dados internacionais, para o relacionamento daquelas já existentes ou para a criação de identificadores internacionais relacionados aos dados de autoridade, encontram, muitas vezes, respaldo em ideais almejadas antes mesmo da existência da Web, tais como a criação de um arquivo de autoridade conjunto a partir do *Linked Systems Project*, desenvolvido nos EUA na década de 1980 (CLACK, 1988, p. 38) e a ideia de um *International Standard Authority Data Number (ISADN)*, que atuaria como um identificador universal desempenhando um papel semelhante ao desempenhado pelo ISBN e o ISSN, ideia esta incorporada à primeira edição das *Guidelines for Authority and References Entries*, publicadas pela IFLA em 1984 (WILLER; DUNSIRE, 2013, p. 8).

Nesta pesquisa, entende-se que as atuais iniciativas para a publicação de dados de autoridade como *Linked Data*, por seu pioneirismo, prestam muito mais do que somente à publicação desses dados. Essas iniciativas prestam-se, principalmente, à exploração inicial do uso de tecnologias da Web Semântica no domínio bibliográfico, à experimentação de diferentes abordagens para a modelagem dos dados, à averiguação da suficiência dos vocabulários, proporcionando o aperfeiçoamento dos já existentes, bem como o desenvolvimento de novos, e à identificação – ou comprovação/contestação – dos benefícios da publicação dos dados de autoridade como *Linked Data*.

As principais tecnologias da Web Semântica foram abordadas no Capítulo 3, seguidas pelos vocabulários de classes e de propriedades, que carregam as possibilidades de modelagem dos dados, assim, convém concluir este capítulo com uma seção acerca dos potenciais benefícios da publicação dos dados de autoridade como *Linked Data*.

4.3 Benefícios dos dados de autoridade publicados como *Linked Data*

Os benefícios e/ou as possibilidades oferecidas pela publicação de *Linked Data* são abordados em diversos momentos na literatura. No domínio bibliográfico, o relatório final do *Library Linked Data Incubator Group*, de 2011, tem se apresentado como a principal síntese desses benefícios. Nesse relatório, os benefícios estão divididos para quatro segmentos intrinsecamente relacionados: (1) pesquisadores, estudantes e usuários, (2) instituições, (3) bibliotecários, arquivistas e curadores, e (4) desenvolvedores e fornecedores de softwares (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Os benefícios elencados no relatório são discutidos nesta seção e complementados com contribuições de outros autores.

Os pesquisadores, os estudantes e os demais usuários dos dados das bibliotecas, para os quais o uso de *Linked Data* pode não estar evidente, podem usufruir de melhores funcionalidades para a descoberta e o uso dos dados e para a navegação entre os dados da biblioteca e além deles. Adicionalmente, os dados das bibliotecas podem ser incluídos de forma estruturada em páginas HTML com o uso do RDFa, o que os torna mais acessíveis às ferramentas de busca gerais, massivamente utilizadas pelos usuários (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Sobre isso, Lampert e Southwick (2013, p. 230, tradução nossa) destacam que o “*Linked Data*, explicado como a transição da Web de documentos para a Web de dados, beneficiará consideravelmente a busca e a descoberta de informações e influenciará as experiências e expectativas dos usuários com as informações contextuais”.

Esses benefícios decorrem principalmente da possibilidade de enriquecimento dos dados a partir de conjuntos de dados externos: “Por meio de *links* ricos com dados complementares de fontes confiáveis, as bibliotecas podem aumentar o valor de seus próprios dados para além da somatória de suas fontes tomadas individualmente” (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, tradução nossa).

Para Gracy, Zeng e Skirvin (2013, p. 20, tradução nossa),

As bibliotecas seriam capazes, a partir de um conjunto de dados, de agregar dados baseados em pedaços/blocos de informações que elas necessitam, sem integrar toda a base de dados ou converter completamente os registros de metadados. Elas podem combinar as declarações de metadados [*metadata statements*] (não registros inteiros) de diferentes *namespaces* ou agregar dados de uma variedade de recursos, baseando-se somente no que é necessário.

No âmbito dos dados de autoridade, o enriquecimento pode ocorrer, por exemplo, com o aproveitamento de formas variantes do nome de uma entidade definidas em outro conjunto de dados, por exemplo, uma instituição pode vincular seus dados aos dados do VIAF e, com isso, utilizar as formas variantes do nome de uma pessoa ou entidade coletiva presentes no VIAF para facilitar ao usuário a localização de recursos informacionais no catálogo da instituição. De forma semelhante, um arquivo de autoridade de assunto pode estar vinculado ao portal *datos.bne.es*, que, por sua vez, está vinculado aos dados da LCSH; partindo desses *links*, o arquivo de autoridade pode ser enriquecido com termos em espanhol e em inglês. Um arquivo de autoridade vinculado ao *Wikidata* pode ser enriquecido, por exemplo, com imagens e informações biográficas de autores.

Relacionada ao enriquecimento dos dados está a integração de dados, entendida como a possibilidade de acessar programaticamente dados de diferentes fontes (MORSHED et al., 2008, p. 37). Morshed et al. (2008, p. 37, tradução nossa) destacam que “Em uma web onde cada fragmento de dado é publicado por meio de tecnologias e formatos de dados padrão, e onde cada fragmento de dado pode ser univocamente nomeado e localizado, a integração de dados [...] é vista como mais próxima agora do que antes”.

Entre os benefícios percebidos pelas instituições, está a possibilidade de redução de custos com as atividades de representação, que demandam conhecimento especializado, uma vez que a publicação como *Linked Data* aumenta as possibilidades de reuso dos dados (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Além disso, por estarem baseados em declarações (triplas) e não em registros, dados de diferentes fontes podem ser integrados para a obtenção dos níveis de granularidade (especificidade) e de exaustividade demandados por determinada instituição.

A publicação dos dados como *Linked Data* e as tecnologias utilizadas para tal podem auxiliar as instituições em seus processos internos de curadoria de dados, bem como na manutenção dos *links* entre seus recursos (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Essa contribuição é observada nos relatos de Vila Suero e Gómez Pérez (2013), sobre o portal *datos.bne.es*, e de Simon et al. (2013), sobre o projeto *data.bnf.fr*:

É importante observar que, embora os catálogos da biblioteca contenham dados de alta qualidade curados por profissionais treinados, ainda há problemas a serem resolvidos. [...] A maior parte dos problemas refletem a evolução do catálogo e são produzidos pelas

mudanças nas regras de catalogação e pela migração de um sistema para outro no decorrer dos anos, entre outros. O processo de geração de LLD [*Library Linked Data*] permitiu detectar de forma semiautomática as deficiências nas fontes de dados. Assim, a curadoria dos dados emerge como um importante valor adicional oferecido pelos processos de LLD. (VILA SUERO; GÓMEZ PÉREZ, 2013, p. 577, tradução nossa).

[...] um dos mais interessantes aspectos da publicação de *Linked Data* fora dos atuais catálogos da biblioteca é o impacto positivo que o princípios do LD, tais como o uso de URIs ao invés de cadeias de caracteres [*strings*] ou o conceito de recursos formalmente tipificados, podem ter na arquitetura de informação das bibliotecas. Exemplos desse impacto benéfico são a inclusão, na BNE, de URIs resolvíveis para recursos equivalentes em conjuntos de dados externos durante o processo de catalogação (por exemplo, a inclusão dos URIs do VIAF e da DBpedia durante a criação de um registro de autoridade) ou a reorganização, em andamento, do catálogo da BNE baseada nas experiências de aplicação dos modelos FR no portal *datos.bne.es*. (VILA SUERO; GÓMEZ PÉREZ, 2013, p. 599, tradução nossa).

Finalmente, combinando as ferramentas da Web Semântica com técnicas de combinação [*matching techniques*], respondemos a grande demanda interna da biblioteca para a melhoria dos catálogos sem a necessidade de aumentar o trabalho do catalogador. Assim, se o *data.bnf.fr* pode conduzir a novos serviços aos usuários externos, ele pode também ser um meio para melhorar e corrigir automaticamente ou semiautomaticamente nossos próprios dados, em longo prazo. Esse é o motivo pelo qual tentamos construir rotinas e mecanismos que podem ser usados dentro dos catálogos originais. Passo a passo, as inferências sobre nossos dados são integradas às fontes originais. Uma interface específica foi desenvolvida, assim os bibliotecários podem validar as combinações obtidas automaticamente. Podemos gerar páginas de obras dentro de nossos arquivos de autoridade ou criar novos *links* entre registros bibliográficos e o arquivo de autoridade de obras, seguindo os princípios do FRBR. O objetivo é reduzir as tarefas de catalogação o máximo possível e criar novos links, de modo que possam ser imediatamente úteis aos usuários finais. (SIMON et al., 2013, p. 14, tradução nossa).

Sobre os benefícios para as bibliotecas, Tharani (2015, p. 5-6, tradução nossa) considera que

Tornar os metadados das bibliotecas coletáveis [*harvestable*] na web pode não somente refinar a precisão e a revocação, mas tem o potencial de empoderar as bibliotecas para que aproveitem o alcance das ferramentas de busca comerciais para continuamente descobrir, navegar e obter novos recursos de conhecimento de domínios específicos [*domain specific knowledge resources*] a partir de seus metadados verificados. Essa é uma ideia romântica e viável, mas sua implementação requer que as bibliotecas saiam de suas zonas de conforto e reforcem o desafio de encontrar soluções colaborativas para ligar as ilhas de informação que temos criado na web para nossos usuários e para nós mesmos.

As rotinas de trabalho dos bibliotecários, arquivistas e curadores de dados são impactadas pelos benefícios apresentados aos usuários e às instituições (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Por exemplo, o reuso e o enriquecimento de dados demandam a alteração de tarefas, bem como a criação de novas e a exclusão daquelas não mais necessárias. Segundo o Relatório final, “Catalogadores serão capazes de concentrar seu esforço em dominar o conhecimento local, ao invés de ter que recriar descrições existentes que já foram elaboradas por outros.” (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, tradução nossa).

Lampert e Southwick (2013, p. 230, tradução nossa) consideram que “o *Linked Data* impactará também nos futuros fluxos de trabalho das bibliotecas e requererá novas perspectivas sobre a ingestão, o gerenciamento, a distribuição e a agregação de grandes conjuntos de dados criados a partir dos metadados das bibliotecas.” Com isso, será necessária uma “mudança de pensamento”:

Os catalogadores precisam pensar no modo como os dados estão modelados quando prosseguem para a catalogação no dia-a-dia; a web semântica começa com o humano criador de dados facilitando o relacionamento enquanto registra os dados. Treinamento e conscientização são ferramentas absolutamente necessárias para fazer as coisas funcionarem adequadamente. (SANTOS; MANCHADO; VILA SUERO, 2015, p. 17, tradução nossa).

Os desenvolvedores e fornecedores de softwares, por sua vez, serão diretamente beneficiados por não estarem restritos aos formatos específicos no domínio bibliográfico. Uma vez que o *Linked Data* tem em sua base modelos e padrões de uso geral na Web,

Os fornecedores de softwares para as bibliotecas que aderirem ao *Linked Data* serão capazes de comercializar seus produtos fora do mundo das bibliotecas, enquanto que fornecedores fora desse mundo poderão ser capazes de adaptar seus produtos mais genéricos para os requisitos específicos das bibliotecas. Com o fomento ao RDF e ao HTTP, os desenvolvedores da comunidade de bibliotecas estão livres da necessidade de utilizar softwares específicos desse domínio, abrindo uma gama crescente de ferramentas genéricas, muitas das quais são *open source*. [...] Isso também abre uma comunidade de desenvolvimento muito maior para prover suporte aos profissionais de tecnologia da informação nas bibliotecas. (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, tradução nossa).

Além dos Formatos MARC 21 e derivados, são específicos do domínio bibliográfico o protocolo Z39.50 e a norma ISO 2709, sendo substituídos, no âmbito das práticas de *Linked Data*, pelo protocolo HTTP e pelos diversos formatos de serialização

dos dados (XML, JSON-LD, *Turtle*, etc.), de utilização não restrita ao domínio bibliográfico (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011). Essas mudanças, além de beneficiarem os desenvolvedores e distribuidores de softwares, facilitarão também “novos e inesperados usos de nossos dados em diferentes comunidades” (SIMON et al., 2013, p. 12, tradução nossa).

Além dos benefícios elencados para os quatro segmentos, destaca-se também que a publicação dos dados das bibliotecas como *Linked Data* apresenta benefícios que vão além do domínio bibliográfico, uma vez que, como apontam Gracy, Zeng e Skirvin (2013, p. 21), ao tornarem seus dados compartilháveis, extensíveis e reutilizáveis, os bibliotecários podem ampliar o alcance de seus catálogos e bases de dados, tornando-se também contribuidores em um espaço global de informação.

Para Catarino e Souza (2012, p. 89),

Transformar os dados bibliográficos em dados ligados não enriquecerá apenas os serviços ofertados pelas bibliotecas aos seus usuários, mas, inclusive, a própria Web, na medida em que será possível, por meio dos mecanismos de busca, acessar os dados bibliográficos presentes nos catálogos de milhares de bibliotecas do mundo todo.

Sobre essa contribuição para a Web, Willer e Dunsire (2013, p. 138-139, tradução nossa) destacam que

Em um ambiente onde “qualquer um pode dizer qualquer coisa sobre qualquer coisa” e máquinas produzem metadados indexando o conteúdo de recursos bibliográficos ou pela análise estatística dos metadados produzidos por humanos, os metadados criados por catalogadores profissionais treinados usando padrões internacionais provavelmente são mais precisos, coerentes, consistentes e completos.

Nesse cenário, os dados de autoridade provenientes do domínio bibliográfico contribuem para a melhoria da recuperação da informação na Web: “Os dados de autoridade das bibliotecas para nomes e assuntos ajudarão a reduzir a redundância de descrições bibliográficas na Web identificando claramente as principais entidades compartilhadas através do *Linked Data*.” (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, tradução nossa). É nesse sentido que Pattuelli (2012, p. 559, tradução nossa) aponta que “Nos últimos anos, muitos profissionais da informação e bibliotecários têm solicitado a implementação de LOD em vocabulários controlados, argumentando que esse é um passo essencial para que se mantenha essa rica informação relevante no ambiente de LOD”.

Sobre a importância dos arquivos de autoridade, van der Graaf e Waaijers (2014, p. 4, tradução nossa) destacam que

Os arquivos de autoridade servem para identificar unicamente “coisas” ou entidades do mundo real, como documentos, pessoas, organizações, e suas propriedades, como relacionamentos e características. Já importantes no mundo clássico das bibliotecas, os arquivos de autoridade são indispensáveis para a recuperação e a análise adequada da informação na era dos computadores. Isso ocorre porque, mais ainda que os humanos, os computadores são ruins ao lidar com a ambiguidade. Por meio de arquivos de autoridade, as pessoas podem dizer aos computadores quais termos, nomes ou números referem-se à mesma coisa ou têm o mesmo significado [...].

Para Pattuelli (2012, p. 559, tradução nossa),

O papel dos vocabulários controlados, incluindo os arquivos de autoridade, ainda está para ser completamente explorado no novo contexto de LOD, em que novas funções como a interligação e a integração de dados tornam-se fundamentais para a descoberta e a navegação.

Por fim, em uma apresentação sobre o *Library of Congress Linked Data Service*, Guenther (2011) sintetiza as razões para o desenvolvimento de um serviço *web* de disponibilização de conjuntos de dados representando vocabulários controlados, razões estas compreendidas aqui também como benefícios da publicação de dados de autoridade como *Linked Data*:

Expor os vocabulários desenvolvidos na LC para uma comunidade mais ampla.
Tornar as listas controladas disponíveis abertamente.
Prover informação abrangente sobre os termos controlados.
Experimentar as tecnologias da *web* semântica e o *linked data*.
Facilitar o processo de desenvolvimento e de manutenção dos vocabulários. (GUENTHER, 2011, slide 15).

Neste capítulo foram apresentados alguns vocabulários para o uso na publicação de dados de autoridade como *Linked Data* (SKOS, MADS/RDF e *RDA Element Sets*) e iniciativas para a publicação desses dados (*Library of Congress Linked Data Service*, *datos.bne.es*, *data.bnf.fr*, VIAF e AGROVOC), bem como uma síntese de seus potenciais benefícios, tanto no domínio bibliográfico quanto além dele. Assim, considerando a exposição acerca dos dados de autoridade (Capítulo 2), acerca do conceito de *Linked Data* e das tecnologias da Web Semântica relacionadas a ele (Capítulo 3) e acerca dos vocabulários e iniciativas (Capítulo 4), apresenta-se, no Capítulo 5, o modelo para a publicação de dados de autoridade como *Linked Data* proposto nesta pesquisa.

5 Modelo para a publicação de dados de autoridade como *Linked Data*

Após a apresentação desta pesquisa, realizada no Capítulo 1, abordou-se o controle de autoridade, os conceitos a ele relacionados e o desenvolvimento dos dados de autoridade nos catálogos em livros, em fichas e digitais (Capítulo 2). No capítulo seguinte, foram apresentados o conceito de *Linked Data* e as principais tecnologias da Web Semântica relacionadas a ele (Capítulo 3), o que levou ao Capítulo 4, em que, a partir da apresentação de vocabulários e de iniciativas, obteve-se um panorama atual acerca da publicação de dados de autoridade como *Linked Data*. O percurso traçado nesses capítulos alinha-se aos apontamentos de Harper e Tillett (2007, p. 53, tradução nossa):

Os benefícios e os méritos do controle de autoridade têm sido debatidos e reafirmados por décadas. Quando aplicamos controle de autoridade, somos lembrados de como ele traz precisão às buscas, como a estrutura sindética de remissivas permite a navegação e provê explicações para variações e inconsistências, como as formas controladas de nomes, títulos e assuntos ajudam a dispor obras durante sua exibição, como podemos realmente relacionar formas autorizadas de nomes, títulos e assuntos utilizadas em várias ferramentas, como diretórios, bibliografias e serviços de resumos e indexação, e assim por diante. Podemos usar a capacidade de relacionamento para incluir os catálogos de bibliotecas em uma mescla das variadas ferramentas disponíveis na Web. Para alcançarmos essa capacidade em um ambiente baseado na Web, será necessário nos movermos de registros que facilitam o relacionamento e a agregação na biblioteca para uma codificação mais geral e universal.

Ainda que nos últimos anos tenham sido desenvolvidos ou adaptados vocabulários de classes e de propriedades para a publicação de dados catalográficos como *Linked Data* e iniciativas ao redor do planeta tenham surgido com esse objetivo, concorda-se com Cole et al. (2013, p. 166-167, tradução nossa): “No contexto das bibliotecas, os serviços baseados em LOD voltados aos usuários finais e em produção ainda são poucos, de maturidade limitada e relativamente não testados”.

Entende-se que esse estágio de experimentação de diferentes abordagens para a publicação de dados como *Linked Data* assemelha-se, em determinados pontos, ao desenvolvimento da automação de bibliotecas, a partir das décadas de 1960 e 1970: a demanda pelo uso de novas tecnologias levou à experimentação de diferentes abordagens e, então, à consolidação daquelas que se mostraram mais adequadas ao

longo dos anos. Willer e Dunsire (2013, p. 178) traçam um paralelo entre a situação atual do uso das tecnologias da Web Semântica e o desenvolvimento da própria Web:

A situação atual assemelha-se ao estágio inicial do desenvolvimento da World Wide Web, antes da disponibilidade de softwares para auxiliar os criadores de páginas web. Os arquivos HTML para os primeiros websites foram criados manualmente utilizando softwares simples de edição de texto [...] Conforme a demanda cresceu e o custo dos dispositivos continuou baixando, foram desenvolvidos softwares melhores para a criação de arquivos HTML. Hoje, a maior parte do HTML é gerada a partir de bancos de dados em tempo real. Os dados são criados em interfaces gráficas amigáveis aos usuários, e esse pode ser um cenário futuro para a Web Semântica. (WILLER; DUNSIRE, 2013, p. 178, tradução nossa).

Na literatura internacional são encontrados estudos que visam a identificar procedimentos relacionados à publicação de *Linked Data*. Esses estudos, ainda que não abordem especificamente os dados de autoridade e suas especificidades, estão aqui indicados por proverem parte do embasamento necessário à apresentação do modelo.

Vila Suero e Gómez Pérez (2013) apresentam o software MARiMbA, projetado para transformar registros nos Formatos MARC 21 em *Linked Data*, e sua utilização na publicação dos dados bibliográficos e de autoridade da BNE no projeto *datos.bne.es*. Para a publicação desses dados, os autores descrevem metas e ações, sintetizadas no Quadro 5: especificação, curadoria dos dados, modelagem, geração, relacionamento, publicação e exploração.

Quadro 5 – Metas e ações do projeto *datos.bne.es*

	1 Especificação	2 Curadoria dos dados	3 Modelagem	4 Geração	5 Relacionamento	6 Publicação	7 Exploração
Meta	Analisar e descrever as características dos dados (fontes de dados e os dados em RDF).	Reparar e melhorar as fontes de dados e o RDF.	Criar um vocabulário para descrever os recursos em RDF.	Produzir recursos RDF a partir das fontes de dados.	Conectar o conjunto de dados em RDF a outros conjuntos de dados relevantes.	Tornar o conjunto de dados disponível na Web.	Definir e desenvolver aplicações que usem o conjunto de dados em RDF.
Ação	1. Identificar e analisar as fontes de dados. 2. Projetar os URIs. 3. Definir uma licença e as informações de proveniência.	1. Curadoria das fontes de dados. 2. Curadoria dos dados em RDF.	1. Analisar e selecionar vocabulários do domínio. 2. Desenvolver o vocabulário. 3. Vocabulário para representar as informações de proveniência.	1. Selecionar, expandir ou desenvolver tecnologias para a produção do RDF. 2. Criar mapeamentos entre o vocabulário e as fontes de dados. 3. Transformar as fontes de dados em RDF.	1. Selecionar os conjuntos de dados alvo para relacionar as entidades do conjunto de dados. 2. Descobrir links com os conjuntos de dados alvo. 3. Validar os links.	1. Publicar o conjunto de dados. 2. Publicar metadados descrevendo o conjunto de dados. 3. Permitir a descoberta efetiva do conjunto de dados.	1. Desenvolver ou configurar aplicações sobre o conjunto de dados.

Fonte: Vila Suero e Gómez Pérez (2013, p. 580, tradução nossa).

Lampert e Southwick (2013) descrevem o *University of Nevada, Las Vegas (UNLV) Linked Data Project*, que teve por objetivo publicar os metadados das coleções digitais da universidade (aproximadamente 45 mil itens, incluindo fotografias, mapas, manuscritos, desenhos arquitetônicos, itens multimídia e histórias orais) como *Linked Data*. O projeto incluiu cinco etapas, sintetizadas no Quadro 6: preparação, transformação, publicação e integração, consumo e avaliação.

Quadro 6 – Visão geral das fases do *UNLV Linked Data Project*

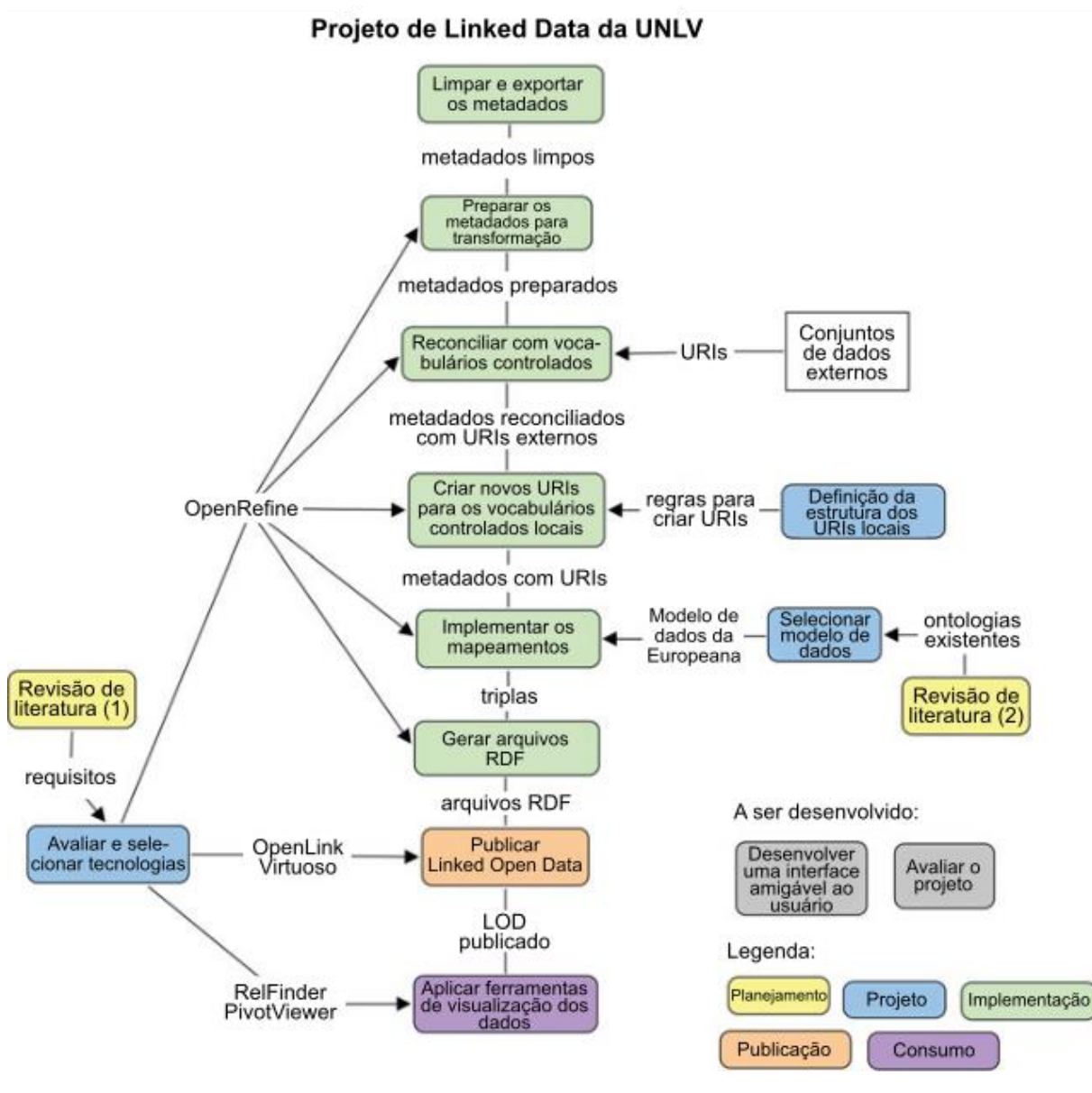
Fase	Entrada	Processos	Saída
Preparação	- Literatura	- Revisão das melhores práticas para os modelos de <i>Linked Data</i> . - Revisão das tecnologias para transformação e gestão de arquivos RDF (<i>triplestore</i>). - Mapeamento entre o esquema de metadados e modelo de dados. - Revisão das melhores práticas para a criação de URIs.	- Modelo de dados. - Escolha das tecnologias. - Mapeamento. - Regras para a criação de URIs.
Transformação	- Modelo de dados. - Tecnologia para apoiar a transformação. - Mapeamento. - Regras para a criação de novos URIs. - Metadados de coleções digitais. - Vocabulários controlados localmente.	- Extração. - Preparação do metadados. - Implementação do mapeamento. - Atribuição de URIs. - Expressão dos vocabulários controlados como <i>Linked Data</i> .	- Arquivos RDF contendo triplas derivadas dos metadados. - Arquivos RDF contendo triplas derivadas de vocabulários controlados locais.
Publicação e integração	- Sistema <i>triplestore</i> . - Outros conjuntos de dados publicados, como a DBpedia.	- Upload dos arquivos RDF em sistema <i>triplestore</i> . - Identificação de dados externos relevantes para serem relacionados. - Criação de triplas contendo os links externos. - Publicação da descrição do conjunto de dados.	- Conjunto de dados da UNLV publicado como <i>Linked Data</i> .
Consumo	- Melhores práticas para sites de <i>Linked Data</i> .	- Projetar e desenvolver a interface do usuário.	- Interface de acesso ao conjunto de dados da UNLV e a outros dados relevantes.

Fase	Entrada	Processos	Saída
Avaliação	- Conjunto de dados da UNLV. - Coleções digitais da UNLV.	- Avaliação das tecnologias, dos benefícios do Linked Data aos usuários e fluxos de trabalho.	- Relatório com recomendações.

Fonte: Lampert e Southwick (2013, p. 245, tradução nossa).

Após esse trabalho inicial, Southwick (2015) rerepresentou o *UNLV Linked Data Project*, dessa vez compreendendo as etapas de planejamento, projeto, implementação, publicação de LOD, e consumo de LOD. Essas etapas estão ilustradas na Figura 24.

Figura 24 – *UNLV Linked Data Project*



Fonte: Southwick (2015, p. 13, tradução nossa).

Além desses trabalhos que não se relacionam aos dados de autoridade, mas que contribuem para a apresentação do modelo, destaca-se também a recomendação publicada pelo W3C *Data on the Web Best Practices* (WORLD WIDE WEB CONSORTIUM, 2017, tradução nossa), que descreve 35 melhores práticas para a publicação de dados na Web, listadas no Quadro 7.

Quadro 7 – Melhores práticas para os dados na Web

<ol style="list-style-type: none"> 1. Forneça metadados 2. Forneça metadados descritivos 3. Forneça metadados estruturais 4. Forneça informações sobre a proveniência dos dados 5. Forneça informações sobre a licença dos dados 6. Forneça informações sobre a qualidade dos dados 7. Forneça um indicador da versão 8. Forneça o histórico de versões 9. Use URIs persistentes como identificadores dos conjuntos de dados 10. Use URIs persistentes como identificadores dentro dos conjuntos de dados 11. Atribua URIs às versões e séries dos conjuntos de dados 12. Use formatos de dados legíveis por máquinas padronizados 13. Use representações dos dados neutras no que se refere ao local 14. Forneça os dados em múltiplos formatos 15. Reuse vocabulários, preferencialmente aqueles padronizados 16. Escolha o nível de formalização correto 17. Forneça o <i>download</i> do conjunto completo 18. Forneça o <i>download</i> de subconjuntos para grandes conjuntos de dados 	<ol style="list-style-type: none"> 19. Use a negociação de conteúdo para oferecer dados disponíveis em múltiplos formatos 20. Forneça acesso em tempo real 21. Forneça dados atualizados 22. Forneça uma explicação para os dados que não estão disponíveis 23. Torne os dados acessíveis por meio de API 24. Use padrões da Web como base para as APIs 25. Forneça a documentação completa para a sua API 26. Evite alterações que quebrem sua API 27. Preserve identificadores 28. Avalie a cobertura do conjunto de dados 29. Obtenha o feedback dos consumidores dos dados 30. Torne o feedback disponível 31. Enriqueça os dados pela geração de novos dados 32. Forneça apresentações complementares 33. Forneça feedback para o publicador original 34. Siga os termos do licenciamento 35. Cita a publicação original
--	--

Fonte: *World Wide Web Consortium* (2017, tradução nossa).

Ainda que publicações como essas orientem sobre algumas práticas e sobre o emprego de algumas tecnologias na publicação de *Linked Data*, nota-se ainda uma carência de orientações que abranjam os dados do domínio bibliográfico, como apontam Cole et al. (2013, p. 193) e Chen (2017, p. 123).

Como resposta a essa demanda por “orientações”, na Figura 25 é apresentado o modelo de publicação de dados de autoridade como *Linked Data* compreendendo políticas, procedimentos e tecnologias, proposto nesta tese a partir do conceito das

funções dos dados de autoridade nos catálogos e de seus benefícios nos ambientes de *Linked Data*.

Figura 25 – Modelo para a publicação de dados de autoridade como *Linked Data*



Fonte: Elaborada pelo autor.

Essas etapas de planejamento; modelagem e mapeamento; tratamento, relacionamento e conversão; publicação; e *feedback* e retroalimentação representadas na Figura 25 são descritas nas seções seguintes.

5.1 Planejamento

A etapa de planejamento compreende o levantamento e as decisões iniciais acerca dos dados de autoridade existentes na instituição e do conjunto de dados que se deseja publicar como *Linked Data*. Os principais procedimentos dessa etapa são:

- seleção e análise dos conjuntos de dados que serão disponibilizados (Seção 5.1.1);

- escolha da licença sob a qual os dados estarão disponíveis (Seção 5.1.2);
- escolha dos conjuntos de dados aos quais os dados poderão estar relacionados (Seção 5.1.3).

5.1.1 Seleção e análise dos conjuntos de dados

O planejamento/elaboração do projeto requer, primeiramente, a seleção do conjunto de dados de autoridade que será publicado como *Linked Data*. Para essa seleção, devem ser respondidas questões como:

- todos os dados de autoridade da instituição serão publicados como *Linked Data* ou somente aqueles relativos a pessoas, entidades coletivas ou conceitos?;
- há algum procedimento de controle de qualidade ou de revisão dos registros que influenciará a escolha de quais dados serão publicados? Por exemplo, serão publicados somente os dados dos registros revisados, por serem considerados dentro do padrão da instituição.

Após a seleção do conjunto de dados, procede-se com sua análise, que tem por objetivo levantar as principais características do conjunto de dados. Em seu relato sobre o *datos.bne.es*, Vila Suero e Gómez Pérez (2013, p. 581) sintetizam a análise com uma apresentação do propósito, do modelo de dados, dos principais conceitos (classes ou entidades representadas pelos dados), dos componentes dos registros de autoridade (por exemplo, campos, subcampos e indicadores, no caso do Formato MARC 21 para Dados de Autoridade), dos detalhes de implementação (formatos de serialização/codificação dos registros de autoridade) e dos identificadores únicos dos dados de autoridade.

Durante a análise, devem ser descritos também os vínculos existentes entre dados de autoridade e dados bibliográficos, se a instituição possui dados bibliográficos explicitamente associados aos dados de autoridade ou se os dados de autoridade são tratados de forma isolada, como pode acontecer no caso de tesouros e outras linguagens documentais. Deve ser levado em consideração se há o interesse da instituição em publicar também os dados bibliográficos, simultânea ou posteriormente, como *Linked Data*.

Essa análise será subsídio para que a equipe envolvida no projeto familiarize-se com os dados de autoridade, obtendo o conhecimento necessário acerca de sua função e

dos padrões utilizados em sua representação (padrões de metadados e padrões de codificação).

5.1.2 Escolha da licença

A escolha das condições sob as quais os dados de autoridade serão disponibilizados e poderão ser (ou não) reutilizados externamente constitui-se, dentro do modelo proposto, uma questão mais política do que técnica, uma vez que perpassa os próprios dados e suas características técnicas, alcançando políticas institucionais e questões jurídicas.

No âmbito político, a escolha de uma licença para a disponibilização dos dados relaciona-se à política institucional de acesso à informação, devendo a licença escolhida estar em acordo com esta política. Nesse sentido, instituições privadas e instituições públicas podem ter comportamentos distintos.

No âmbito jurídico, a escolha de uma licença leva em conta a natureza dos dados e a legislação vigente, que pode determinar os possíveis direitos sobre eles. Assim, um conjunto de dados, tais como os dados de autoridade, pode não ser considerado, por determinada legislação, propriedade intelectual da mesma forma que uma obra literária, uma imagem ou uma música, no entanto, outros direitos de propriedade podem recair sobre ele, levando em conta o investimento demandado para a coleta, o tratamento e a apresentação dos dados, conforme relatado por (RODRÍGUEZ DONCEL; GÓMEZ PÉREZ; MIHINDUKULASOORIYA, 2013, p. 3-4).

No Brasil, os direitos sobre uma base de dados estão descritos na Lei nº 9.610, de 19 de fevereiro de 1998, a lei dos direitos autorais:

Art. 87. O titular do direito patrimonial sobre uma base de dados terá o direito exclusivo, a respeito da forma de expressão da estrutura da referida base, de autorizar ou proibir:

I - sua reprodução total ou parcial, por qualquer meio ou processo;

II - sua tradução, adaptação, reordenação ou qualquer outra modificação;

III - a distribuição do original ou cópias da base de dados ou a sua comunicação ao público;

IV - a reprodução, distribuição ou comunicação ao público dos resultados das operações mencionadas no inciso II deste artigo. (BRASIL, 1998).

Ainda que os catálogos e os arquivos de autoridade possam enquadrar-se nesta lei, na literatura brasileira não foram encontrados estudos ou relatos sobre a aplicação de licenças a dados bibliográficos ou de autoridade, estejam publicados como *Linked Data* ou não. Para Stephens (2011, p. 247, tradução nossa), “os termos de uso dos dados obtidos em catálogos de biblioteca frequentemente não são claros [...]”. É desse mesmo ano o relatório final do *Library Linked Data Incubator Group*, que sintetiza as questões envolvidas no licenciamento dos dados das bibliotecas, principalmente dos catálogos: os dados podem ser restritos às políticas locais, contratos e condições, variando de país para país, sendo a questão de sua propriedade complicada pelo compartilhamento dos dados entre bibliotecas nos últimos 50 anos, que envolve cópia, modificação, melhoria e agregação em catálogos coletivos regionais, nacionais e internacionais (W3C LIBRARY LINKED DATA INCUBATOR GROUP, 2011, item 3.3).

Observando esse cenário em que os dados dos catálogos ainda careçam de licenças e de discussões acerca de sua propriedade, entende-se aqui que a publicação dos dados de autoridade como *Linked Data* é uma oportunidade para as instituições considerarem a necessidade de atribuir uma licença aos seus dados – preferencialmente uma licença aberta, seguindo os princípios de *Linked Open Data* e seu papel enquanto defensoras do livre conhecimento na sociedade – favorecendo, assim, o compartilhamento, o relacionamento e o reuso de seus dados na Web. É neste sentido que caminha a *Open Knowledge International*, com a publicação de uma breve declaração listando os *Principles on Open Bibliographic Data*:

Ao publicar dados bibliográficos, faça uma declaração de licença explícita e robusta. [...]

Utilize uma renúncia ou licença apropriadas aos dados. [...]

Se você deseja que seus dados sejam definitivamente usados e adicionados a outros, eles devem ser abertos como definido na Open Definition (<http://opendefinition.org/>) – em particular, cláusulas para uso não comercial e outras restrições não devem ser usadas. [...]

Quando possível, recomendamos a disponibilização explícita dos dados bibliográficos em domínio público via PDDL ou CC0. (OPEN KNOWLEDGE FOUNDATION, 2011, p. 2, tradução nossa).

Embora estes princípios façam menção apenas aos dados bibliográficos, no âmbito da declaração, os dados de autoridade são considerados dados secundários inclusos no conceito de dados bibliográficos (OPEN KNOWLEDGE INTERNATIONAL, 2011). A *Open Definition* mencionada na declaração aponta princípios que definem a abertura em

relação aos dados e ao conteúdo, mais especificamente, o que é compreendido como uma obra⁸⁸ aberta (“*open work*”) e quais os requisitos para que uma licença seja considerada uma licença aberta (“*open license*”) (OPEN KNOWLEDGE FOUNDATION, 2015). A compreensão de obra aberta apresentada nessa definição pode ser aplicada aos dados de autoridade:

Uma **obra** aberta deve satisfazer os seguintes requisitos em sua distribuição:

1.1 Licença ou status aberto

A **obra** *deve* estar em **domínio público** ou sob uma **licença aberta** [...] Quaisquer termos adicionais que acompanhem a obra (tais como termos de uso ou patentes mantidas pelo licenciador) *não devem* contradizer o status de domínio público da obra ou os termos de sua licença.

1.2 Acesso

A **obra** *deve* ser fornecida como um todo [...], e *deve* ser descarregável via Internet sem cobrança. Qualquer informação necessária para a observância da licença (tal como os nomes dos contribuidores) também *deve* acompanhar a obra.

1.3 Legibilidade por máquinas

A **obra** *deve* ser fornecida em uma forma prontamente processável por um computador e na qual os elementos individuais da obra possam ser facilmente acessados e modificados.

1.4 Formato aberto

A **obra** *deve* ser fornecida em um formato aberto. Um formato aberto é aquele que não coloca restrições, monetárias ou não, sobre o seu uso e pode ser completamente processado pelo menos em uma ferramenta free/libre/open source. (OPEN KNOWLEDGE INTERNATIONAL, 2015, grifo do autor, tradução nossa).

As licenças utilizadas para a publicação de *Linked Data* são classificadas de acordo com suas liberdades e restrições, das mais abertas às mais fechadas. No Quadro 8 é apresentada a classificação utilizada por Rodríguez Doncel, Gómez Pérez e Mihindikulasooriya (2013, p. 8).

Quadro 8 – Classificação das licenças a partir de suas restrições

Tipo de licença	Licença
Domínio público	<i>Creative Commons</i> 1.0 Universal (CC0 1.0) – Dedicção ao Domínio Público <i>Open Data Commons – Public Domain Dedication and Licence</i> (ODC-PDDL)

⁸⁸ Na *Open Definition* “O termo obra sera utilizado para denotar o item ou parte de conhecimento que está sendo transferida”.

Tipo de licença	Licença
Atribuição	<i>Creative Commons – Attribution</i> (CC-BY) <i>Open Data Commons – Attribution</i> (ODC-BY)
Compartilhamento pela mesma licença	<i>Creative Commons – Attribution – ShareAlike</i> (CC-BY-SA) <i>Open Data Commons – Open Database License</i> (ODC-ODbL) <i>United Kingdom – Open Government Licence</i> (UK-OGL) <i>GNU Free Documentation License</i> (GFDL)
Com restrições	<i>Creative Commons – Attribution – NonCommercial</i> (CC-BY-NC) <i>Creative Commons – Attribution – NoDerivatives</i> (CC-BY-ND) <i>Creative Commons – Attribution – NonCommercial – NoDerivatives</i> (CC-NC-ND), etc.
Fechadas	Todos os direitos reservados, etc.



Fonte: Adaptado de Rodríguez Doncel, Gómez Pérez e Mihindukulasooriya (2013, p. 8).






Além dessas licenças, outras são criadas para atender demandas específicas de determinada instituição ou governo, como é o caso da licença *French Open Licence*, utilizada no *data.bnf.fr*, criada a partir da *Creative Commons Attribution License*, embora o uso de uma licença já padronizada seja mais benéfico:

Além dos benefícios da melhoria da eficiência organizacional e da economia de custos, o uso de termos de licenciamento padrão pode levar a uma maior interoperabilidade dos dados bem como ao aumento da consciência do usuário sobre os termos da licença, permitindo, assim, sua melhor observância. (KORN; OPPENHEIM, 2011, p. 4, tradução nossa).

Para a Korn e Oppenheim (2011, p. 4), o uso de licenças *Creative Commons* (CC) em conjuntos de dados tem como vantagem a ampla utilização dessas licenças, sua familiaridade, flexibilidade e disponibilidade em formato legível para humanos e para máquinas. As licenças CC incluem sete licenças, apresentadas no Quadro 9.

Quadro 9 – Licenças *Creative Commons*

Ícone da licença	Síntese
	CC0 1.0 Universal - Dedicção ao Domínio Público (CC0 1.0) Dedicção ao domínio público, com renúncia a todos os seus direitos sob as leis de direito de autor e/ou de direitos conexos referentes à obra, em todo o mundo, na medida permitida por lei. É possível copiar, modificar, distribuir e executar a obra, mesmo para fins comerciais, tudo sem pedir autorização.
	Atribuição (CC BY) Permite que outros distribuam, remixem, adaptem e criem a partir de sua obra, mesmo para fins comerciais, desde que lhe atribuam o devido crédito pela criação original

	<p>Atribuição, Compartilhamento pela mesma licença (CC BY-SA)</p> <p>Esta licença permite que outros remixem, adaptem e criem a partir da sua obra, mesmo para fins comerciais, desde que lhe atribuam o devido crédito e que licenciem as novas criações sob termos idênticos.</p>
	<p>Atribuição, Sem derivações (CC BY-ND)</p> <p>Esta licença permite a redistribuição, comercial e não comercial, desde que a obra seja distribuída inalterada e no seu todo, com crédito atribuído a você.</p>
	<p>Atribuição, Não comercial (CC BY-NC)</p> <p>Esta licença permite que outros remixem, adaptem e criem a partir da sua obra para fins não comerciais e, embora as novas obras tenham de lhe atribuir o devido crédito e não possam ser usadas para fins comerciais, os usuários não têm de licenciar essas obras derivadas sob os mesmos termos.</p>
	<p>Atribuição, Não comercial, Compartilhamento pela mesma licença (CC BY-NC-SA)</p> <p>Esta licença permite que outros remixem, adaptem e criem a partir da sua obra para fins não comerciais, desde que atribuam a você o devido crédito e que licenciem as novas criações sob termos idênticos.</p>
	<p>Atribuição, Não comercial, Sem derivados (CC BY-NC-ND)</p> <p>Esta é a mais restritiva das nossas seis licenças principais, só permitindo que outros façam download dos seus trabalhos e os compartilhem desde que atribuam crédito a você, mas sem que possam alterá-los de nenhuma forma ou utilizá-los para fins comerciais.</p>

Fonte: Elaborado pelo autor a partir de Creative Commons (2017).

Entre as licenças CC, a CC0 é ideal para a publicação de *Linked Open Data* em domínio público, uma vez que implica a renúncia de todos os direitos sobre os dados. Outra licença apropriada à publicação de *Linked Open Data* em domínio público é a *Open Data Commons – Public Domain Dedication and Licence* (ODC-PDDL), que permite a cópia, a distribuição, o uso, a produção de obras a partir da base de dados, a modificação, a transformação e a criação de novas bases de dados sem qualquer restrição (OPEN KNOWLEDGE INTERNATIONAL, 2008).

Há também a *Open Data Commons – Attribution License* (ODC-By), que concede as mesmas liberdades da ODC-PDDL, no entanto, com a condição de que o devido crédito seja dado à base de dados original (OPEN KNOWLEDGE INTERNATIONAL, 2010), e a *Open Data Commons – Open Database License* (ODbL), que inclui as liberdades e condições da ODC-By, além de requerer que as bases de dados ou obras derivadas sejam licenciadas também pela ODC-By e que a base de dados, redistribuída ou adaptada, seja mantida aberta (OPEN KNOWLEDGE INTERNATIONAL, 2014).

No que diz respeito ao *Linked Data*, Rodríguez Doncel, Gómez Pérez e Mihindikulasooriya (2013, p. 4-5) apontam que a declaração de direitos deve considerar de forma independente os seguintes níveis: uma única tripla; um conjunto de dados ou grafo RDF; os links relacionando conjuntos de dados, os chamados mapeamentos; e os recursos externos referenciados nos dados. Apesar da existência desses três níveis, os autores concluem em seu estudo que as declarações de direitos têm sido aplicadas de modo desigual entre eles, sendo que as declarações concentram-se nos conjuntos de dados (RODRÍGUEZ DONCEL; GÓMEZ PÉREZ; MIHINDUKULASOORIYA, 2013, p. 11).

No modelo de dados RDF, o conjunto de dados também é considerado um recurso, sendo identificado por um URI, podendo, assim, ser descrito por meio de declarações. A descrição de um conjunto de dados inclui informações como título, criador, data, etc. e licença. Nas declarações que descrevem a licença de um conjunto de dados podem ser utilizadas propriedades dos vocabulários *Dublin Core*, *Creative Commons Rights Expression Language*⁸⁹, *Open Data Rights Statement Vocabulary*⁹⁰, *Licenses for Linked Open Data*⁹¹ e *Linked Data Rights*⁹². Apesar dessa variedade de vocabulários, “por enquanto, somente *Creative Commons RDF ccREL* tem sido utilizado pela comunidade de *Linked Data*” (RODRÍGUEZ DONCEL; GÓMEZ PÉREZ; MIHINDUKULASOORIYA, 2013, p. 8-9). Como valores, essas propriedades podem receber URIs que identifiquem as licenças. As declarações acerca das licenças serão incluídas na descrição do conjunto de dados, realizada na etapa de publicação dos dados, abordada na Seção 5.4.1.

A publicação de dados sob licenças abertas, ainda que não seja um pré-requisito para o *Linked Data*, tem se mostrado uma tendência que recebe cada vez mais aderência, como se observa pelo movimento de *Linked Open Data*, cujo crescimento é ilustrado pela quantidade de conjuntos de dados representados no *Linking Open Data cloud diagram*, apresentado no capítulo 3. Destaca-se também que todas as iniciativas descritas na Seção 4.2 disponibilizam seus dados sob licenças abertas: domínio público, *Creative Commons Public Domain Dedication (CC0)*, *French Open Licence*, *Open Data Commons*

⁸⁹ Disponível em: <<https://creativecommons.org/ns>>. Acesso em: 12 dez. 2017.

⁹⁰ Disponível em: <<http://schema.theodi.org/odrs/>>. Acesso em: 12 dez. 2017.

⁹¹ Disponível em: <http://ns.inria.fr/14lod/v2/14lod_v2.html>. Acesso em: 12 dez. 2017.

⁹² Disponível em: <<http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/>>. Acesso em: 12 dez. 2017.

Attribution License (ODC-BY) e *Creative Commons Attribution 3.0 Unported License*, o que mostra a aderência ao movimento de LOD no âmbito dos dados de autoridade.

5.1.3 Escolha dos conjuntos de dados para relacionamento

Partindo da seleção e análise preliminar dos dados, é possível identificar potenciais conjuntos de dados para o relacionamento com os dados de autoridade que serão publicados. Esses conjuntos de dados são internos (da mesma instituição) ou externos (de outras instituições), já publicados como *Linked Data* ou não. A identificação inicial dos potenciais conjuntos é importante, pois, a partir dela, os procedimentos seguintes (modelagem, escolha dos vocabulários, mapeamentos, etc.) podem ser aprimorados para assegurar a implantação, atual ou futura, desses relacionamentos.

Os conjuntos de dados internos podem incluir outros conjuntos de dados de autoridade e conjuntos de dados bibliográficos, disponíveis em arquivos de autoridade, catálogos, bibliotecas digitais, repositórios institucionais, etc. Ainda que esses conjuntos de dados não estejam publicados como *Linked Data*, a instituição pode antecipar-se e, a partir da publicação dos dados de autoridade, rever seus conjuntos e suas práticas, no intuito de preparar-se para a futura publicação desses dados como *Linked Data*.

Os conjuntos de dados externos incluem arquivos de autoridade de outras instituições, tais como agências bibliográficas nacionais, que contenham registros para as mesmas pessoas, entidades coletivas, conceitos, etc. representados no conjunto de dados que será publicado. O levantamento dos conjuntos de dados externos é feito a partir das políticas e manuais de controle de autoridade e dos dados existentes nos próprios registros de autoridade, por exemplo:

- uma instituição que utiliza o Formato MARC 21 para Dados de Autoridade, ao incluir em seu arquivo de autoridade um registro criado por outra instituição, mantém o identificador do registro original no campo 001 – “Número de controle”, 010 – “Número de controle da LC”, 016 – “Número de controle de uma agência bibliográfica nacional”, ou 035 – “Número de controle em um sistema”, junto com dados que identificam a instituição de origem do registro; os dados presentes nesses campos indicam possíveis conjuntos de dados para o relacionamento;

- por considerar a LC uma referência internacional para registros de autoridade, uma instituição pode utilizá-la como fonte de consulta e indicar em seus registros o ponto de acesso autorizado como definido na LC, o mesmo pode ocorrer com o arquivo de autoridade de outra biblioteca nacional ou com linguagens documentais como o Descritores em Ciências da Saúde (DeCS)⁹³ e *Medical Subject Headings* (MeSH)⁹⁴; essa política da instituição já indica possíveis conjuntos de dados externos para o relacionamento.

Além de arquivos de autoridade externos, são identificados também conjuntos de dados criados fora do domínio bibliográfico, mas que estão relacionados aos dados de autoridade, por exemplo, DBpedia, Wikidata, Geonames, ORCID⁹⁵ e instalações do software VIVO⁹⁶. A importância dos relacionamentos com dados de outros domínios é destacada por Marcondes (2017, p. 3, destaque do autor) ao referir-se aos acervos de bibliotecas, arquivos e museus:

Mais que simplesmente publicar esses conteúdos, é necessário dar-lhes visibilidade e integrá-los aos fluxos de informação da Web, tanto criando “links” entre acervos complementares como fazendo com que sejam criados “links” para esses acervos desde outras páginas fora do domínio específico das instituições de memória e cultura.

A identificação dos potenciais conjuntos de dados para o relacionamento com os dados de autoridade publicados como *Linked Data*, conduzida durante o planejamento/elaboração do projeto, constitui-se como um levantamento inicial, podendo conjuntos de dados ser acrescentados ou removidos posteriormente, em razão da existência ou não de condições para que os relacionamentos sejam representados em declarações RDF, conforme é abordado na Seção 5.3.2.

Além dos procedimentos descritos nessa etapa, também fazem parte do planejamento a definição da equipe responsável e do cronograma de execução do projeto com os recursos financeiros disponíveis e necessários. Indica-se também a inclusão de ações para a conscientização dos profissionais criadores e utilizadores dos dados de autoridade que não estiverem na equipe responsável pelo projeto. Essa conscientização demanda, entre outros, a apresentação dos benefícios da publicação dos

⁹³ Disponível em: <<http://decs.bvs.br/>>. Acesso em: 12 dez. 2017.

⁹⁴ Disponível em: <<https://www.nlm.nih.gov/mesh/>>. Acesso em: 12 dez. 2017.

⁹⁵ Disponível em: <<https://orcid.org/>>. Acesso em: 12 dez. 2017.

⁹⁶ Disponível em: <<http://vivoweb.org/>>. Acesso em: 12 dez. 2017.

dados de autoridade como *Linked Data*, de modo a alcançar a adesão dos profissionais aos objetivos do projeto.

5.2 Modelagem e mapeamento dos dados

Concluído o planejamento inicial, procede-se com a modelagem e o mapeamento dos dados, que tem por objetivo principal a definição de um modelo, que guiará a organização do conjunto de dados, e sua expressão utilizando classes e propriedades em RDF. Para tanto, inclui as seguintes etapas que, embora apresentadas nesta sequência, são em diversos momentos realizadas de forma concomitante:

- modelagem dos dados (Seção 5.2.1);
- mapeamento entre os dados de autoridade, o modelo de dados e as classes e propriedades utilizadas em sua descrição (Seção 5.2.2);
- descrição do vocabulário utilizado (Seção 5.2.3);
- definição de padrões para a construção dos URIs (Seção 5.2.4).

5.2.1 Modelagem dos dados

Modelos de dados, no sentido amplo utilizado na Seção 2.3, são encontrados em vários contextos, por exemplo, o modo com que os dados estão dispostos e relacionados em catálogos em livros, em fichas ou digitais constitui um modelo de dados, assim como se pode afirmar que os Formatos MARC 21 possuem um modelo de dados próprio, consistindo este do conjunto de dados presentes em registros MARC 21 e suas inter-relações. Nas palavras de Serra e Santarem Segundo (2017, p. 174), “o modelo de dados é a forma onde o relacionamento entre recursos e seus metadados são documentados”.

O RDF é um modelo de dados que permite a descrição de recursos por meio de declarações (triplas) compostas por um recurso, uma propriedade e um valor. O RDF, no entanto, não é o único modelo de dados utilizado na publicação de *Linked Data*: tendo como base o RDF, são definidos modelos de dados para representar os recursos de determinado domínio. Nesses modelos de dados, são estabelecidas as classes e as propriedades utilizadas na descrição dos recursos.

Van Hooland e Verborgh (2015, p. 12, tradução nossa) entendem que, para a disponibilização de recursos e de seus metadados de forma estruturada na Web, deve-se

decidir quais características são as mais importantes para serem representadas e, ao se fazer isso, tem-se “uma abstração da realidade a partir do desenvolvimento de um modelo”.

Para a publicação de *Linked Data*, uma instituição pode optar por reutilizar um modelo de dados existente ou criar ou adaptar um para atender suas especificidades. Esse modelo de dados guiará a organização, em um sentido amplo, do conjunto de dados que será publicado pela instituição. Os três vocabulários apresentados na Seção 4.1 – SKOS, MADS/RDF e *RDA Element Sets* – possuem seus próprios modelos de dados, os quais, mesmo abrangendo o mesmo domínio, contêm suas particularidades. As iniciativas apresentadas na Seção 4.2 – *LC Linked Data Service*, *datos.bne.es*, *data.bnf.fr*, VIAF e AGROVOC – utilizam modelos de dados de algum desses vocabulários ou próprios.

A escolha, a criação ou a adaptação de um modelo deve levar em conta sua interoperabilidade semântica com outros modelos, utilizados em outros vocabulários e conjuntos de dados. A falta dessa interoperabilidade semântica advém da heterogeneidade semântica, que “[...] ocorre no significado das classes e propriedades do modelo devido a diferentes interpretações das entidades do mundo real [...]” (ZAPOUNIDOU; SFAKAKIS; PAPTAEODOROU, 2017, p. 546, tradução nossa).

A modelagem dos dados de autoridade em RDF requer o entendimento de que:

- os dados de autoridade estão, fora do ambiente de *Linked Data*, agrupados em registros de autoridade;
- os registros de autoridade estão agrupados em arquivos de autoridade, sendo que um arquivo de autoridade pode dar origem a um ou mais conjuntos de dados ou passar a fazer parte de outros conjuntos;
- um registro de autoridade refere-se a um objeto do mundo real, por exemplo, uma pessoa, uma instituição, um lugar ou um conceito;
- os dados de autoridade presentes em um registro podem se referir ao objeto do mundo real, ao próprio registro ou, ainda, a outros dados de autoridade, por exemplo:
 - o ano de nascimento de uma pessoa é um dado que se refere ao objeto do mundo real;
 - a data de criação é um dado que se refere ao registro de autoridade;

- o idioma ou alfabeto de um termo preferido é um dado que se refere a outro dado, o próprio termo.

A distinção entre um objeto ou entidade do mundo real e a sua representação tem sido discutida em áreas como a Filosofia, a Linguística, a Semiótica, a Inteligência Artificial e a Ciência da Informação, com diversos e distintos aportes. Na publicação de *Linked Data*, essa questão influencia a modelagem dos dados e as decisões sobre a estrutura dos URIs (Seção 5.2.4).

Na iniciativa de *Linked Data* da BNE, *datos.bne.es*, tem-se, por exemplo, o URI <http://datos.bne.es/resource/XX1029037>, que identifica o objeto do mundo real Machado de Assis, declarado como indivíduo da classe pessoa. Esse URI é utilizado como recurso em declarações que descrevem os pontos de acesso autorizado e variantes para o autor, suas datas de nascimento e morte e as fontes consultadas para a criação de seu registro de autoridade. O recurso Machado de Assis é também declarado como equivalente ao recurso <http://id.loc.gov/authorities/names/n80002329>, do *Linked Data Service* da LC, como apresentado na Figura 26.

Figura 26 – Machado de Assis na *datos.bne.es*

```
@prefix ns2: <http://datos.bne.es/def/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://datos.bne.es/resource/XX1029037> rdf:type ns2:C1005 .
<http://datos.bne.es/resource/XX1029037> rdfs:label "Machado de Assis" .
<http://datos.bne.es/resource/XX1029037> ns2:P5002 "1839-1908" .
<http://datos.bne.es/resource/XX1029037> ns2:P1002 "Quincas Borba, 1979" .
<http://datos.bne.es/resource/XX1029037> ns2:P1002 "WWW LC auth. file, 4-1-2001" .
<http://datos.bne.es/resource/XX1029037> ns2:P1002 "Dom Casmurro, 1987" .
<http://datos.bne.es/resource/XX1029037> owl:sameAs <http://id.loc.gov/authorities/names/n80002329> .
```

Fonte: Adaptado de *datos.bne.es*.

No *Linked Data Service* da LC, para a representação de pessoas e de entidades coletivas, emprega-se um modelo de dados proveniente do MADS/RDF e um do SKOS. No modelo de dados proveniente do MADS/RDF, por exemplo, o URI <http://id.loc.gov/authorities/names/n80002329> é utilizado como recurso em declarações que descrevem os pontos de acesso autorizado e variantes para Machado de Assis, as fontes consultadas para a criação de seu registro de autoridade e os metadados

administrativos desse registro, tais como as datas da criação e da última alteração, a agência catalogadora e o idioma da catalogação. Ainda nesse modelo, é declarado que <http://id.loc.gov/authorities/names/n80002329> identifica o objeto do mundo real <http://id.loc.gov/rwo/agents/n80002329>. Essa declaração é feita com a propriedade *madsrdf:identifiesRWO*, que “[...] relaciona um recurso representando um conceito de alguma coisa à coisa em si, ao objeto do mundo real” (LIBRARY OF CONGRESS, 2015, tradução nossa):

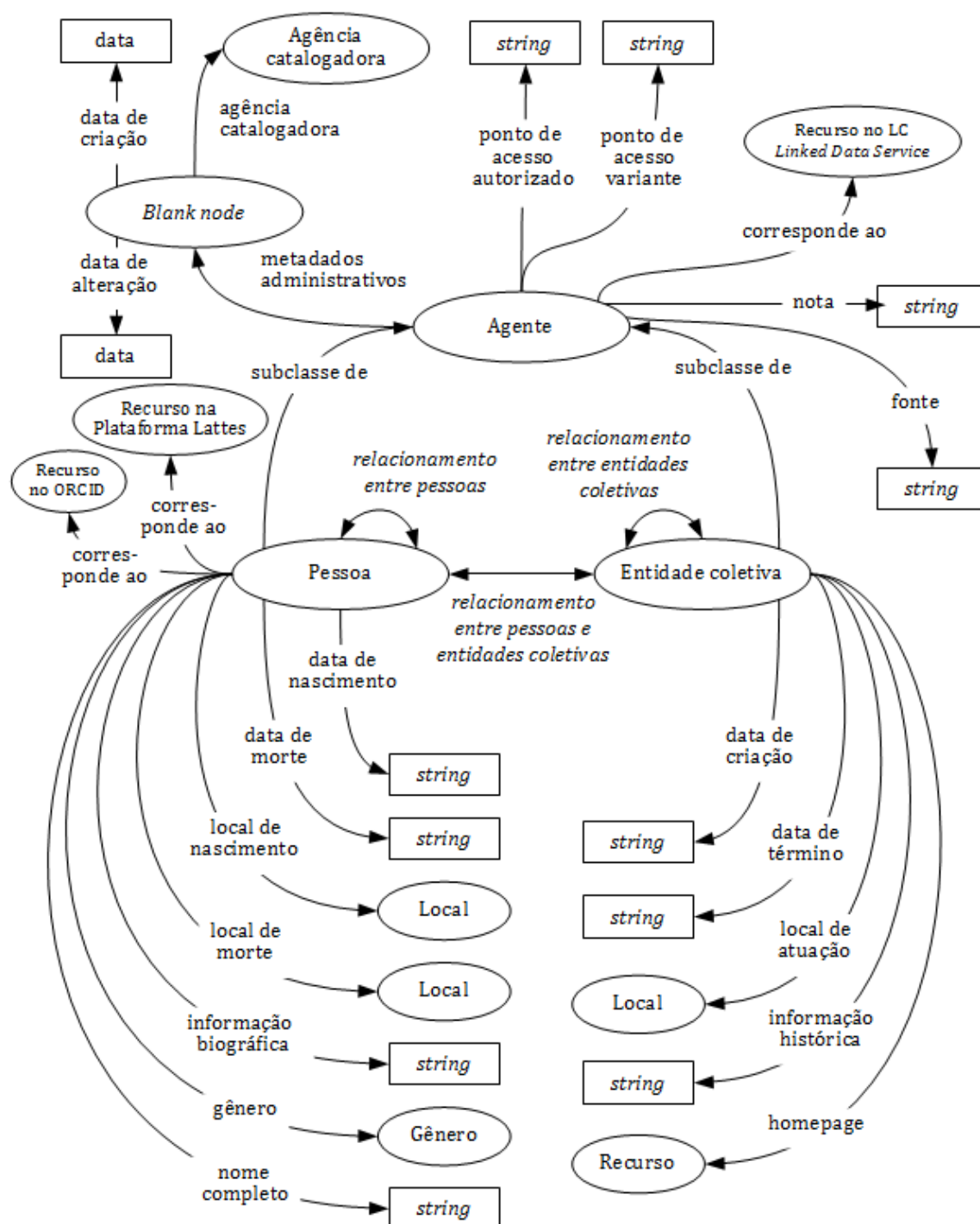
```
<http://id.loc.gov/authorities/names/n80002329> madsrdf:identifiesRWO
    <http://id.loc.gov/rwo/agents/n80002329>.
```

No modelo de dados proveniente do SKOS, o mesmo URI <http://id.loc.gov/authorities/names/n80002329> é utilizado para identificar Machado de Assis, no entanto, não há menção ao objeto do mundo real identificado por <http://id.loc.gov/rwo/agents/n80002329>.

Para a modelagem dos dados, é necessário, além do conhecimento específico do domínio e dos modelos existentes, o conhecimento das possibilidades do RDFS e da OWL, que serão utilizadas no momento de expressar o modelo de dados por meio de classes e propriedades. Nesse ponto, faz-se um alerta quanto ao uso das classes e propriedades da OWL, que, por destinarem-se às ontologias, que demandam nível de expressividade maior que o demandado pelo *Linked Data*, podem trazer para o ambiente de *Linked Data* uma complexidade não desejada, dificultando o reuso dos dados.

A título de exemplificação, a Figura 27 apresenta o esboço de um modelo de dados para representar dados de autoridade.

Figura 27 – Modelo de dados para dados de autoridade



Fonte: Elaborada pelo autor.

Algumas questões que podem auxiliar na modelagem dos dados de autoridade são:

- quais tipos de entidades estão representadas nos registros de autoridade?

- quais dados estão presentes nos registros de autoridade?
- desses dados, quais se referem à entidade representada pelo registro? E quais se referem ao próprio registro?
- existem dados utilizados localmente apenas para as rotinas de controle de autoridade e que não devem ser publicados como *Linked Data*, tais como a identificação dos responsáveis pela criação e pelas alterações do registro ou a indicação de que o registro foi ou não revisto dentro de um processo de controle de qualidade?
- quais dados relacionam os registros de autoridade a registros de outros arquivos de autoridade?
- quais dados pretende-se incluir aos registros de autoridade a partir de conjuntos de dados externos?

Após a modelagem dos dados, procede-se com a escolha das classes e propriedades que serão utilizadas para descrever os dados de autoridade, essa escolha é realizada durante o mapeamento, apresentado na seção seguinte.

5.2.2 Mapeamento

A etapa de mapeamento objetiva, tendo como intermediário o modelo de dados, a identificação de correspondências entre os dados de autoridade, disponíveis no arquivo de autoridade da instituição e selecionados para serem publicados como *Linked Data*, e as classes e as propriedades dos vocabulários que serão utilizados na publicação desses dados. Esse mapeamento assemelha-se àquele já estudado na literatura da Ciência da Informação e relacionado aos padrões de metadados:

O processo de estabelecer os relacionamentos entre metadados semanticamente equivalentes em diferentes padrões é denominado na literatura como mapeamento (*mapping* ou *crosswalking*). O resultado do mapeamento é chamado de mapa (*map* ou *crosswalk*) e consiste em uma representação visual dos relacionamentos, das equivalências e das lacunas entre os padrões de metadados mapeados. (ASSUMPÇÃO; SANTOS, 2014, p. 3811, grifo do autor).

O mapeamento entre padrões de metadados com o objetivo de converter ou transformar registros, como o abordado por Assumpção e Santos (2014), ainda que tenha seus objetivos e conceitos próprios, apresenta semelhanças e diferenças em relação ao mapeamento que tem como alvo classes e propriedades de vocabulários em

RDF. Uma das diferenças é que, no âmbito da publicação de *Linked Data*, o mapeamento está inerente à definição, à redefinição e ao refinamento das propriedades e elementos de vocabulários (DUNSIRE et al., 2011, p. 30).

Para a conversão, o mapeamento é sempre de um padrão de metadados de origem para padrão de destino, estando presente o conceito de “registro”: um conjunto de dados seguindo determinado padrão de metadados. Por outro lado, em se tratando de classes e propriedades, diferentes vocabulários RDF podem ser utilizados ao mesmo tempo em um mapeamento, visto que não há o conceito de registro, sendo cada declaração (tripla) considerada individualmente.

Como semelhança, tem-se a necessidade de considerar os diferentes níveis de granularidade ou de detalhes em questão, visto que os vocabulários podem apresentar classes e propriedades com diferentes níveis de detalhamento entre si e em relação aos dados da instituição. Uma instituição define em seus dados, por exemplo, o tipo entidade coletiva, para diferenciar entidades coletivas de pessoas; há, portanto, no modelo de dados, a classe “Entidade coletiva”. A instituição define também que a “Entidade coletiva” tem como subclasse a classe “Conferência”, que abrange conferências, congressos, simpósios, etc. Durante o mapeamento para as classes de um vocabulário, nota-se que ele possui apenas uma classe equivalente à classe “Entidade coletiva”, sem qualquer subclasse equivalente a “Conferência”. Até este ponto, tem-se um problema que também ocorre no mapeamento entre padrões de metadados: a falta de equivalências entre os padrões mapeados (ASSUMPÇÃO; SANTOS, 2014, p. 3812). No entanto, são as possíveis soluções para esse problema que também diferem o “tradicional” mapeamento entre padrões de metadados do mapeamento entre classes e propriedades de vocabulários. Como soluções para esse problema, têm-se:

- o uso de uma classe de outro vocabulário como equivalente à classe “Conferência”, uma vez que é possível utilizar diferentes vocabulários em um mesmo conjunto de dados; ou
- a criação de uma classe “Conferência”, como parte de um vocabulário criado e mantido pela própria instituição.

A primeira solução, que envolve o reuso de um vocabulário já existente, é a mais recomendada, enquadrando-se entre as melhores práticas publicadas pelo W3C (2017, tradução nossa): “Reuse vocabulários, preferencialmente aqueles padronizados”.

A utilização de vocabulários já em uso por outros capta e facilita consenso nas comunidades. Isso aumenta a interoperabilidade e reduz redundâncias, encorajando, desse modo, o reuso de seus dados. De forma específica, o uso de vocabulários compartilhados para metadados (especialmente aqueles estruturais, de proveniência, de qualidade e de versionamento) ajuda na comparação e no processamento automático dos dados e metadados. Além disso, o uso de códigos e de termos de padrões ajuda a reduzir a ambiguidade e os conflitos entre elementos ou valores similares. (WORLD WIDE WEB CONSORTIUM, 2017, tradução nossa).

Sobre a possibilidade de uso simultâneo de classes e de propriedades de diferentes vocabulários em um mesmo conjunto de dados, adverte-se aqui quanto aos cuidados necessários ao respeito às classes e às propriedades como definidas em seus vocabulários de origem, por exemplo, o respeito aos *ranges* e domínios de uma propriedade, de modo que os dados publicados pela instituição possam ser reutilizados e reunidos com outros conjuntos de dados.

Um exemplo de mapeamento envolvendo classes e propriedades de diferentes vocabulários e criadas pela instituição é apresentado no Quadro 10, elaborada a partir da Figura 27.

Quadro 10 – Exemplo de mapeamento

Campo e subcampos dos registros de autoridade	Representação no modelo de dados	Propriedades dos vocabulários utilizados
1XX – Cabeçalho	Ponto de acesso autorizado	<i>skos:prefLabel</i>
4XX – Remissiva “Ver”	Ponto de acesso variante	<i>skos:altLabel</i>
670 – Fonte positiva de dados	Fonte consultada	<i>rdaz:P60001</i> “has source consulted”
667 – Nota geral não pública 680 – Nota geral pública	Nota	<i>rdaz:P60002</i> “has cataloguer’s note”
100\$q – Forma completa do nome 378 – Forma completa do nome	Nome completo	<i>rdaa:P50115</i> “has fuller form of name”
375 – Gênero	Gênero	<i>rdaa:P50116</i> “has gender”
100\$d – Datas associadas à pessoa 046\$f – Data de nascimento	Data de nascimento	<i>rdaa:P50121</i> “has date of birth”
100\$d – Datas associadas à pessoa 046\$g – Data de morte	Data de morte	<i>rdaa:P50120</i> “has date of death”
370\$a – Local de nascimento	Local de nascimento	<i>rdaa:P50119</i> “has place of birth”

Campo e subcampos dos registros de autoridade	Representação no modelo de dados	Propriedades dos vocabulários utilizados
370\$b – Local de morte	Local de morte	<i>rdaa:P50118</i> <i>"has place of death"</i>
678\$a\$b – Dados biográficos ou históricos	Informação biográfica	<i>rdaa:P50113</i> <i>"has biographical information"</i>
	Informação histórica	<i>rdaa:P50035</i> <i>"has corporate history"</i>
670\$u – Fonte positiva de dados – URI 856\$u – Localização e acesso eletrônicos – URI	Currículo Lattes	<i>skos:exactMatch</i>
010\$a – Número de controle na LC	Número de controle na LC	
670\$u – Fonte positiva de dados – URI 856\$u – Localização e acesso eletrônicos – URI	ORCID	
046\$q – Data de estabelecimento	Data de criação	<i>rdaa:P50037</i> <i>"has date of establishment"</i>
046\$r – Data de término	Data de término	<i>rdaa:P50038</i> <i>"has date of termination"</i>
500 – Remissiva “Ver também” – Pessoa	Relacionamentos entre pessoas	<i>rdaa:P50106</i> <i>"has real identity"</i> <i>rdaa:P50105</i> <i>"has alternate identity"</i>
510 – Remissiva “Ver também” – Entidade coletiva	Relacionamentos entre entidades coletivas	<i>rdaa:P50008</i> <i>"has hierarchical superior"</i> <i>rdaa:P50010</i> <i>"has hierarchical subordinate"</i> <i>rdaa:P50016</i> <i>"has successor"</i> <i>rdaa:P50012</i> <i>"has predecessor"</i>

Fonte: Elaborado pelo autor.

O mapeamento com vistas à publicação de *Linked Data* vai além da aparente e relativamente simples correspondência de um campo com uma propriedade RDF, como pode parecer pelo Quadro 10, demandando o entendimento do contexto em que cada campo e propriedade se enquadra, de seu relacionamento com outros campos e propriedades e, principalmente, da diferença entre o “paradigma” em que os dados de autoridade atualmente se encontram e o “paradigma” do *Linked Data* e da Web

Semântica. Algumas questões envolvidas nesse entendimento são discutidas por Dunsire et al. (2011) e Chen (2015).

O resultado do mapeamento, apresentado em quadros e/ou diagramas, como o exemplificado no Quadro 10, será utilizado na etapa de conversão dos dados para RDF, apresentada na Seção 5.3.3.

5.2.3 Descrição do vocabulário utilizado

As classes e propriedades reutilizadas de outros vocabulários, junto com aquelas criadas para atender às necessidades específicas dos dados de autoridade de uma instituição, constituem o vocabulário utilizado no projeto.

A descrição desse vocabulário é parte importação da documentação do projeto e pode ser feita em formato voltado aos humanos, em uma página HTML, por exemplo, e para a máquina, em RDF. Sua principal finalidade é tornar explícito o modelo de dados utilizado e as relações entre suas classes e propriedades, e, no âmbito do *Linked Open Data*, tornar essas informações públicas e acessíveis. Exemplos desses vocabulários criados para atender as necessidades de descrição são:

- *BnF-onto*⁹⁷ – criado para a iniciativa *data.bnf.fr*; nesse iniciativa houve a reutilização de classes e propriedades do BIO⁹⁸, DC, FOAF, ISNI⁹⁹, *RDA Element Sets* e SKOS, de modo que a *BnF-onto* inclui somente a descrição nas propriedades que foram criadas especificamente pela BnF; a descrição deste vocabulário está disponível em RDF/XML;
- *Agrontology*¹⁰⁰ – criado para o AGROVOC; como houve a reutilização das classes e propriedades do SKOS e do SKOS-XL, o *Agrontology* inclui somente as propriedades que foram criadas para o AGROVOC; o vocabulário está disponível em uma página HTML e em um arquivo em RDF/XML;
- *BNE ontology*¹⁰¹ – criado para a iniciativa *datos.bne.es*; abrange também dados bibliográficos; inclui classes e propriedades criadas a partir dos vocabulários

⁹⁷ Disponível em: <<http://data.bnf.fr/ontology/bnf-onto/>>. Acesso em: 12 dez. 2017.

⁹⁸ Disponível em: <<http://vocab.org/bio/0.1/>>. Acesso em: 12 dez. 2017.

⁹⁹ Disponível em: <<http://www.isni.org/>>. Acesso em: 12 dez. 2017.

¹⁰⁰ Disponível em: <<http://aims.fao.org/aos/agrontology>>. Acesso em: 12 dez. 2017.

¹⁰¹ Disponível em: <<http://datos.bne.es/def/ontology.html>>. Acesso em: 12 dez. 2017.

FRBR e ISBD, publicados pela IFLA¹⁰², *RDA Element Sets*, DC e Bibo e além de propriedades criadas pela própria BNE (BIBLIOTECA NACIONAL DE ESPAÑA, 2016b); as classes e as propriedades que já existiam em outros vocabulários não foram reaproveitadas, mas sim definidas novamente no vocabulário da BNE; o vocabulário está disponível em uma página HTML e em um arquivo em RDF/XML.

5.2.4 Definição de padrões para os URIs

Os URIs são a base para a publicação de *Linked Data* e sua correta criação é um dos desafios para a utilização prática do RDF pelas comunidades de descrição de recursos, visto o papel que desempenham enquanto recursos, propriedades e valores nas declarações (FERREIRA; SANTOS, 2013, p. 21). Assim, a instituição deve atentar-se aos URIs utilizados para identificar:

- o conjunto de dados que será publicado como *Linked Data*;
- os registros de autoridade que integram o conjunto de dados;
- os objetos do mundo real aos quais os registros de autoridade se referem;
- as classes e as propriedades dos vocabulários externos;
- as classes e as propriedades do vocabulário criado pela instituição;
- os recursos, definidos pela instituição ou externamente, utilizados como valores nas declarações, por exemplo, idiomas, locais e outros valores controlados por listas externas e internas.

A maioria desses URIs será definida pela instituição, enquanto outros, que identificam recursos externos, tais como classes, propriedades e valores, serão definidos por terceiros, cabendo à instituição sua correta utilização.

Como orientação para a definição dos URIs HTTP, tem-se o guia *Cool URIs for the Semantic Web*, publicado pelo W3C, que, em síntese, apresenta duas estratégias: URIs com *hash* e URIs 303 (WORLD WIDE WEB CONSORTIUM, 2008).

Na primeira estratégia, ao final do URI que identifica um documento, é acrescentado um símbolo de *hash* (#) e um fragmento, de modo que o URI com o *hash* e o fragmento passa a identificar um objeto do mundo real, por exemplo:

¹⁰² Disponível em: <<http://origin.iflstandards.info/ns/fr/>>. Acesso em: 12 dez. 2017.

- *http://pessoas.com/12345* identifica o registro de autoridade (recurso informacional) que representa a pessoa Machado de Assis;
- *http://pessoas.com/12345#pessoa* identifica a pessoa (objeto do mundo real) Machado de Assis.

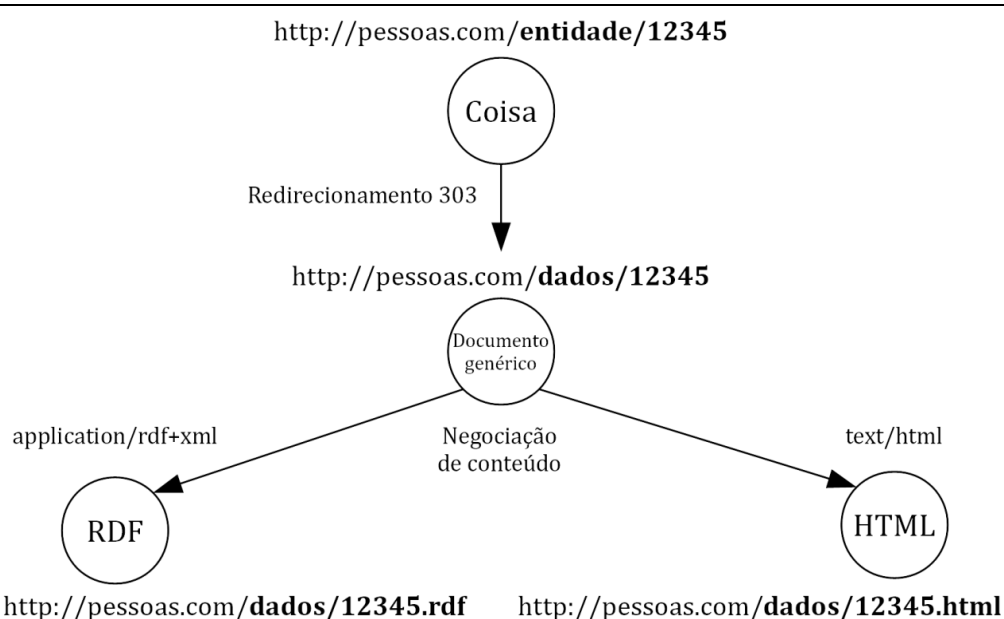
A estratégia dos URIs com *hash* pode ser utilizada também com a negociação de conteúdo, de modo a prover a descrição em diversos formatos de apresentação e/ou serialização (WORLD WIDE WEB CONSORTIUM, 2008).

Na segunda estratégia, o código *303 See Other* do protocolo HTTP, um código de redirecionamento, é utilizado pelo servidor para indicar um documento Web com informações (descrição) do recurso identificado pelo URI. Por exemplo:

- o URI *http://pessoas.com/entidade/12345* identifica a pessoa Machado de Assis;
- o URI *http://pessoas.com/dados/12345* identifica a descrição da pessoa Machado de Assis;
- o servidor, ao receber *http://pessoas.com/entidade/12345*, redireciona para *http://pessoas.com/dados/12345*.

Essa estratégia pode ser utilizada com a negociação de conteúdo, de modo a prover a descrição em diversos formatos de apresentação e/ou serialização, como exemplificado na Figura 28.

Figura 28 – Uso do código 303 com negociação de conteúdo



Na iniciativa *datos.bne.es* emprega-se a estratégia do URI 303 com a negociação de conteúdo, como verificado com a ferramenta *Vafu*¹⁰³:

- o URI *http://datos.bne.es/resource/XX1029037* identifica a pessoa (objeto do mundo real) Machado de Assis;
- os registros de autoridade (recursos informacionais) que descrevem esse objeto do mundo real são identificados e localizados pelos URIs:
 - *http://datos.bne.es/persona/XX1029037.html*
 - *http://datos.bne.es/persona/XX1029037.rdf*
 - *http://datos.bne.es/persona/XX1029037.ttl*.

Essa estratégia é notada também no *Linked Data Service* da LC. Uma vez que cada estratégia tem suas vantagens e desvantagens, o guia aponta que os URIs com *hash* devem ser preferidos para pequenos conjuntos de dados, por exemplo, um vocabulário de classes e propriedades. O guia do W3C lista também outras recomendações relacionadas à simplicidade, estabilidade, gerenciamento, relacionamento de URIs e negociação de conteúdo (WORLD WIDE WEB CONSORTIUM, 2008).

A utilização de identificadores é algo recorrente nas práticas de catalogação: ISBNs identificam publicações monográficas, ISSNs identificam publicações periódicas, códigos de barras e números de tomo identificam exemplares, números de controle identificam registros nos Formatos MARC 21 em programas de catalogação cooperativa e em redes de catalogação, e códigos internos podem ser utilizados para identificar registros em um sistema de gerenciamento de bibliotecas e permitir os vínculos entre eles. Esses identificadores, embora funcionem nos catálogos e arquivos de autoridade, não são adequados enquanto identificadores para a publicação de *Linked Data*, uma vez que não são URIs do tipo HTTP, como preconizam as boas práticas listadas por Berners-Lee (2006). No entanto, esses identificadores já existentes podem ser utilizados como base na criação desses URIs.

A criação de URIs a partir de identificadores já existentes, denominados *natural keys*, é descrita por Dodds e Davis (2012) e consiste no acréscimo do identificador já existente a um URI base. O número de controle de um registro de autoridade criado no Formato MARC 21 é um identificar possível para a criação de URIs, como observado em iniciativas como o *Linked Data Service* e o *datos.bne.es*, que possuem URIs criados a

¹⁰³ Disponível em: <<http://vafu.redlink.io/>>. Acesso em: 12 dez. 2017.

partir desses números de controle. Nos casos em que não for possível a utilização de um identificador já existente, novos identificadores são criados, como acontece no VIAF, em que a publicação dos dados de autoridade gerou novos registros de autoridade.

Em algum momento, pode-se questionar o uso de pontos de acesso autorizados para a composição dos URIs, por exemplo, *http://tesauro.com.br/conceitos/educacao-infantil* para o registro com o ponto de acesso autorizado “Educação infantil” ou *http://pessoas.com/machado-de-assis* para Machado de Assis. Neste trabalho, este uso não é recomendado, visto que pontos de acesso autorizados podem passar por mudanças ao longo da existência dos registros de autoridade e sua utilização comprometeria a persistência dos URIs.

Outras recomendações sobre a definição de URIs estão disponíveis na literatura. Dodds e Davis (2012), por exemplo, descrevem padrões para a criação de URIs e van Hooland e Verborgh (2015) focam nos URIs que também são URLs, apresentando algumas práticas que são adequadas a longo prazo e outras que não o são, tais como a inclusão de fragmentos referentes a tecnologias ou softwares, que, por tornarem-se obsoletos, prejudicam o uso dos URIs.

5.3 Tratamento, relacionamento e conversão dos dados

O mapeamento e os padrões para os URIs, resultados das etapas anteriores, são alguns dos componentes necessários:

- ao tratamento dos dados (Seção 5.3.1);
- ao relacionamento com outros conjuntos de dados (Seção 5.3.2); e
- à conversão dos registros de autoridade, em seu formato de origem, para triplas em RDF (Seção 5.3.3).

5.3.1 Tratamento dos dados

Esta etapa, aqui denominada tratamento dos dados, é descrita na literatura com denominações como curadoria dos dados (*data curation*) (VILA SUERO; GÓMEZ PÉREZ, 2013), limpeza de metadados (*cleaning metadata*) (SOUTHWICK, 2015) ou simplesmente limpeza (*cleaning*) (VAN HOOLAND; VERBORGH, 2015).

Para a Vila Suero e Gómez Pérez (2013, p. 596, tradução nossa), o objetivo é avaliar e assegurar a qualidade das fontes de dados utilizadas e dos dados publicados como *Linked Data*. Van Hooland e Verborgh (2015, p. 88, tradução nossa) também relacionam a limpeza dos dados ao conceito de qualidade de dados e, segundo os autores,

Mesmo com mais e mais pessoas falando sobre a importância da qualidade dos dados em diferentes domínios de aplicação, esse assunto continua difícil de lidar. Primeiramente, qualidade é uma noção intangível. É frequentemente definida como a extensão na qual um produto ou serviço satisfaz uma necessidade. Mas as necessidades mudam, o que torna a qualidade dos dados dependente do contexto de uso. [...] A disponibilização de metadados como linked data está fundamentalmente baseada na premissa de que você permite que outros reuam seus metadados, independentemente do contexto de uso.

Chen (2017, p. 129) entende que a conversão completa de dados de bibliotecas para *Linked Data* não é possível sem a intervenção e a edição humana, mesmo que regras de mapeamento tenham sido definidas claramente.

No que tange aos dados de autoridade, o tratamento dos dados volta-se ao padrão ou formato de origem dos dados com o objetivo de identificar e corrigir eventuais falhas em seu uso e nos valores dos metadados que possam causar erros durante a conversão dos dados para RDF ou durante seu uso, após serem publicados como *Linked Data*. Como componentes necessários a essa etapa, tem-se a documentação relativa ao padrão ou formato e às políticas e aos manuais de catalogação e controle de autoridade da instituição, que ditam o preenchimento dos valores.

Tomando como exemplo registros de autoridade no Formato MARC 21, o tratamento dos dados focaria no uso correto dos campos, subcampos e indicadores desse formato e nos dados preenchidos como valores nesses campos, subcampos e indicadores. Seguindo esse exemplo, durante o tratamento, seriam identificados e corrigidos:

- a ausência de campos, subcampos e indicadores de uso obrigatório;
- o uso de campos, subcampos e indicadores inexistentes ou já obsoletos no Formato MARC 21 para Dados de Autoridade;
- o uso de valores em desacordo com as políticas e instruções da instituição;
- registros de autoridade duplicados;

- registros de autoridade legados de sistemas anteriores ou criados de acordo com políticas não mais vigentes.

Em teoria, a existência de um arquivo de autoridade permite o controle dos pontos de acesso utilizados nos dados bibliográficos e no próprio arquivo, ou seja, um arquivo de autoridade possui pontos de acesso consistentes e, por meio deles, permite a consistência dos pontos de acesso controlados de um catálogo. No entanto, por razões que fogem do escopo desta pesquisa, na prática, arquivos de autoridade podem apresentar inconsistências, tais como a existência de mais de um registro de autoridade, com diferentes pontos de acesso autorizados, para uma mesma entidade, os chamados registros duplicados. Uma vez que podem prejudicar o uso dos dados de autoridade quando publicados como *Linked Data*, essas consistências, se possível, precisam ser corrigidas na etapa de tratamento dos dados.

As formas de correção dos dados de autoridade dependem de sua quantidade e dos recursos humanos e tecnológicos disponíveis. A remoção e/ou a mesclagem de registros de autoridade duplicados, por exemplo, pode ser feita pelo sistema de gerenciamento de bibliotecas, se este dispuser de tal funcionalidade. As correções que não puderem ser realizadas, por exemplo, aquelas em registros legados ou criados seguindo políticas obsoletas, podem demandar a revisão do mapeamento realizado entre o padrão de metadados, o modelo de dados e as classes e propriedades em RDF (Seção 5.2.2), ou, ainda, dependendo da extensão em que tais registros divergem do modelo de dados e das políticas atuais, eles podem demandar procedimentos e softwares de conversão específicos ou serem excluídos do conjunto de registros que será publicado como *Linked Data*.

A utilização de softwares no tratamento dos dados depende dos formatos ou padrões em que os dados se encontram. Nesta seção são descritos brevemente três softwares que, além do tratamento dos dados, podem ser utilizados também na conversão para RDF, abordada na Seção 5.3.3: *MarcEdit*, *Catmandu* e *OpenRefine*.

O *MarcEdit*¹⁰⁴ é um software gratuito voltado a registros nos Formatos MARC 21 que, com suas funcionalidades para o tratamento dos dados, permite:

- alterar, em lote ou um a um, campos, subcampos, indicadores e dados dos registros;

¹⁰⁴ Disponível em: <<http://marcedit.reset.net/>>. Acesso em: 12 dez. 2017.

- gerar relatórios sobre os campos e subcampos utilizados em um conjunto de registros;
- identificar registros duplicados;
- validar registros;
- converter registros nos Formatos MARC 21 codificados de acordo com a ISO 2709 em registros na forma de tabelas e vice-versa.

O *Catmandu*¹⁰⁵ é um software livre definido como “[...] uma ferramenta de linha de comando para acessar e converter dados de biblioteca digital, serviços de pesquisa ou qualquer outro conjunto de dados abertos” (CATMANDU, 2017, tradução nossa). Durante o tratamento dos dados, o Catmandu pode ser utilizado na conversão dos registros de autoridade em diferentes formatos, entre eles, registros nos Formatos MARC 21 codificados com a ISO 2709 em tabelas.

O *OpenRefine*,¹⁰⁶ que também é um software livre, tem seu uso no tratamento de dados descrito por van Hooland e Verborgh (2015). Os autores indicam que, para o uso do *OpenRefine*, os dados devem estar armazenados em tabelas e, para aqueles casos em que a forma de armazenamento não for tabular, por exemplo, um banco de dados relacional ou registros nos Formatos MARC 21, deve-se transformar os dados em um modelo tabular, em que cada coluna corresponda a um metadado e cada linha corresponda a um registro (VAN HOOLAND; VERBORGH, 2015, p. 78). As tabelas contendo os dados são, então, carregadas no *OpenRefine* e, a partir das funcionalidades desse software:

- são verificados se os metadados estão preenchidos com os valores e/ou formatos de valores esperados, por exemplo, se um metadado referente a uma data foi realmente preenchido com uma data e no formato de data apropriado;
- os valores preenchidos em um metadado são separados para dar origem a novos metadados, por exemplo, o nome completo de uma pessoa e suas datas de nascimento e morte, quando constarem em um ponto de acesso autorizado, são extraídos da coluna referente a esse metadado para formar novas colunas na planilha;

¹⁰⁵ Disponível em: <<http://librecat.org/>>. Acesso em: 12 dez. 2017.

¹⁰⁶ Disponível em: <<http://openrefine.org/>>. Acesso em: 12 dez. 2017.

- os valores preenchidos em duas ou mais colunas são reunidos para compor uma única coluna;
- os valores são alterados por meio de funções como localizar e substituir;
- os valores são padronizados com o uso de técnicas de cluster que identificam e substituem as formas variantes de um valor por uma forma escolhida pelo utilizador do *OpenRefine*;
- os registros duplicados são identificados e removidos, entre outras funcionalidades.

Esses três softwares, assim como outros, podem ser utilizados de forma conjunta no tratamento dos dados, visto que possuem algumas funcionalidades que se repetem entre eles e outras que são exclusivas. Para exemplificar sua utilização em conjunto no tratamento de registros de autoridade, é apresentado o seguinte exemplo:

- os registros de autoridade no formato MARC 21 são exportados no sistema de gerenciamento de bibliotecas com a codificação ISO 2709;
- por meio de relatórios de campos e subcampos, gerados com o *MarcEdit*, são identificados campos e subcampos inadequados e a ausência de campos e subcampos obrigatórios;
- os erros são corrigidos no sistema de gerenciamento de bibliotecas e os registros são exportados novamente com a codificação ISO 2709;
- utilizando o *Catmandu* ou o *MarcEdit*, o arquivo com os registros em ISO 2709 é transformado em uma tabela no formato CSV, em que cada coluna é um campo ou subcampo do MARC 21 e cada linha é um registro; os campos e subcampos incluídos nessa tabela são aqueles mapeados para o modelo de dados e as classes e propriedades em RDF (Seção 5.2.2);
- a tabela com os registros é carregada no *OpenRefine* e lá são realizadas verificações e correções no conteúdo dos campos e subcampos, as colunas são mescladas ou divididas, dando origem a novas colunas, e os valores são padronizados, quando necessário.

Após o tratamento, os dados serão relacionados a outros conjuntos de dados. Essa etapa de relacionamento é apresentada na seção seguinte.

5.3.2 Relacionamento com outros conjuntos de dados

No modelo proposto nesta pesquisa, o relacionamento dos dados de autoridade com outros conjuntos de dados é dividido em dois momentos: o relacionamento por meio de *links* já existentes nos dados de autoridade e o relacionamento a partir da descoberta de novos *links*.

Os dados de autoridade, ainda que não estejam publicados como *Linked Data*, podem apresentar *links* para outros arquivos de autoridade. Em registros de autoridade no Formato MARC 21, por exemplo, *links* para outros arquivos de autoridade são encontrados nos campos:

- 010 (Número de controle na LC), 016 (Número de controle na agência bibliográfica nacional) e 035 (Número de controle no sistema), que comportam, exclusivamente, identificadores de registros;
- 7XX, conjunto de campos dedicados aos relacionamentos entre arquivos de autoridade; nesses campos, é possível registrar os identificadores dos registros e/ou o ponto de acesso que é autorizado para representar determinada entidade em outro arquivo de autoridade.

Os *links* entre registros de diferentes arquivos de autoridade, no entanto, para que possam ser publicados como *Linked Data*, devem ser devidamente declarados com o uso de URIs HTTP. Para isso, é necessário que o conjunto de dados alvo, para o qual o *link* apontará, esteja publicado como *Linked Data*. Se um registro de autoridade contiver o identificador do registro ao qual ele está relacionado e tal identificador tiver sido utilizado na construção do URI deste registro, o URI do registro relacionado poderá ser criado automaticamente, a partir de inclusão de um URI base. Por exemplo, o número de controle da LC para o registro de autoridade que representa Machado de Assis é *n 80002329*; se esse número estiver no campo 010 do registro de autoridade, poderá ser utilizado para compor o URI <http://id.loc.gov/authorities/names/n80002329>, bastando, para isso, o acréscimo do URI base <http://id.loc.gov/authorities/names/>.

Do contrário, se o registro de autoridade não contiver o identificador do registro ao qual ele se relaciona, mas possuir algum dado que possa relacioná-lo a tal registro, por exemplo, os dois registros possuem o mesmo ponto de acesso autorizado, a descoberta do *link* e sua indicação com um URI HTTP poderão ser feitas

automaticamente por meio da reconciliação (*reconciling*) no *OpenRefine* (HANSON, 2014, p. 15; VAN HOOLAND; VERBORGH, 2015).

Van Hooland e Verborgh (2015, p. 134) apresentam a reconciliação como a verificação, de forma automática, de um valor para identificar se ele consta em um conjunto de dados externo. No caso dos dados de autoridade, os pontos de acesso autorizado e variantes são utilizados nessa verificação. No *OpenRefine*, a reconciliação é realizada com a extensão *RDF Refine*,¹⁰⁷ que compara os dados que se deseja relacionar com um conjunto de dados presente em um arquivo RDF (*RDF dump*); ou com os dados obtidos a partir da consulta, em tempo real, a um SPARQL *endpoint* (HANSON, 2014, p. 15; CARLSON; SEELY, 2017). Por exemplo, os dados de autoridade que representam pessoas e entidades coletivas podem ser comparados com um arquivo contendo os dados do VIAF.

Nesse processo de relacionamento de conjuntos de dados, em se tratando especificamente de tesouros, Morshed et al. (2011, p. 37) destacam como desafios as diferenças nos formatos, na estrutura, na semântica e nos rótulos dos conceitos em diferentes idiomas.

Além dos links apontando para outros conjuntos de dados criados a partir de arquivos de autoridade representando pessoas, entidades coletivas, conceitos, etc., é possível também estabelecer *links* para conjuntos de dados que representam outros valores utilizados nos dados de autoridade, tais como idiomas e locais. O estabelecimento desses links pode ocorrer por meio da criação de URIs, a partir de seus identificadores, ou pela reconciliação.

Para evitar relacionamentos inadequados, tem-se, ainda, uma etapa de validação ou revisão dos *links* (VILA SUERO; GÓMEZ PÉREZ, 2013, p. 596), principalmente daqueles descobertos a partir da reconciliação. Considerando a quantidade de dados e de links, essa validação pode ocorrer por amostragem ou não. Para Bojārs, Žogla e Eglīte (2015, p. 185), erros podem ser encontrados durante a descoberta de *links* ou durante a revisão em busca de falsos positivos e falsos negativos, de modo que

[...] o algoritmo de identificação de correspondências pode ser melhorado para que lide com os erros ou alerte os usuários sobre os erros que devem ser corrigidos. Como resultado, o relacionamento de conjuntos de dados é um processo iterativo em que os resultados das execuções iniciais do algoritmo são usados para melhorar as execuções seguintes e a qualidade geral dos conjuntos de dados envolvidos. (BOJĀRS; ŽOGLA; EGLĪTE, 2015, p. 185, tradução nossa).

¹⁰⁷ Disponível em: <<http://refine.deri.ie/>>. Acesso em: 12 dez. 2017.

Dada uma das principais características do *Linked Data* – o relacionamento entre os dados de forma processável por softwares –, a descoberta de *links* entre os dados de autoridade de uma instituição e conjuntos de dados externos permite, ainda, a descoberta de novos *links*, para outros conjuntos de dados (BOJĀRS; ŽOGLA; EGLĪTE, 2015, p. 187). E esses novos *links* podem ser utilizados no enriquecimento dos dados de autoridade da instituição (AGENJO; HERNÁNDEZ; VIEDMA, 2011; VAN HOOLAND; VERBORGH, 2015). Para Bojārs, Žogla e Eglīte (2015, p. 187, tradução nossa),

O enriquecimento dos dados é uma tarefa complexa e os detalhes de como pode ser realizada dependem dos conjuntos de dados envolvidos. Por exemplo, os campos dos registros podem ser copiados de um conjunto de dados para o outro, convertidos e mesclados quando necessário. No caso de taxonomias, tais como os dados de autoridade de bibliotecas, os registros relacionados de ambas as taxonomias podem conter rótulos em diferentes idiomas e esses registros podem ser enriquecidos a partir da cópia dos rótulos entre os conjuntos de dados, facilitando a criação de taxonomias multilíngues.

A partir de URIs do VIAF, por exemplo, uma instituição pode obter URIs para outros conjuntos de dados e reutilizar os dados desses conjuntos para enriquecer seus dados de autoridade, seja incluindo novos dados (datas, pontos de acesso autorizado e variantes em outros idiomas ou alfabetos, imagens, etc.), seja possibilitando novos *links* para a navegação de seus usuários.

5.3.3 Conversão para RDF

A etapa de conversão visa à criação de arquivos em RDF a partir do mapeamento entre os dados de autoridade existentes, o modelo de dados e as classes e as propriedades em RDF. São necessários a essa etapa o mapeamento (Seção 5.2.2) e a documentação relativa ao padrão de metadados em que os dados de autoridade se encontram, aos vocabulários em RDF, ao formato de serialização e às tecnologias que serão utilizadas para a conversão dos dados.

Diferentes tecnologias têm sido utilizadas para converter os dados em triplas RDF. A utilização de folhas de estilo criadas com a XSLT, por exemplo, é relatada por Malmsten (2009), Hanson (2014) e Cole et al. (2017). Summers et al. (2008) relatam o uso da linguagem *Python* por meio da biblioteca *pymarc*¹⁰⁸. Van Hooland e Verborgh

¹⁰⁸ Disponível em: <<https://pypi.python.org/pypi/pymarc>>. Acesso em: 12 dez. 2017.

(2015) descrevem o uso do *OpenRefine*, com a extensão *RDF Refine*, e Simon et al. (2013) apresentam o *CubicWeb*, utilizado na iniciativa *data.bnf.fr*. Já Vila Suero, Villazón Terrazas e Gómez Pérez (2012) descrevem o uso do software MARiMba, criado no âmbito da iniciativa *datos.bne.es*.

Entende-se aqui que a escolha das tecnologias utilizadas é influenciada principalmente por fatores como a disponibilidade de pessoal com conhecimento das tecnologias cogitadas e o acesso ou não a elas, tecnologias gratuitas e/ou livres, por exemplo, podem ter preferência se comparadas às pagas e/ou fechadas.

O resultado da conversão será um ou mais arquivos em RDF, em um formato de serialização escolhido pela instituição, que serão utilizados na publicação dos dados (Seção 5.4).

Arquivos de autoridade, assim como catálogos bibliográficos, geralmente estão em constante processo de modificação, seja pela inclusão de novos registros, seja pela alteração ou exclusão daqueles já existentes, como aponta Clack (1990, p. 6-7, tradução nossa):

Os dados em arquivos de autoridade não permanecem estáticos. Eles são dinâmicos e estão em um constante estado de mudança. Eles mudam por uma ampla variedade de razões. Quando o conteúdo de um acervo muda, expande ou contrai, os arquivos de autoridade são afetados. Quando o código de catalogação muda, ou é revisto ou reinterpretado, o arquivo de autoridade deve mudar para refletir o novo código. Também, quando autores e entidades coletivas responsáveis mudam seus nomes, mudanças serão necessárias no arquivo de autoridade.

Alterações no modelo de dados e nos vocabulários utilizados também demandam atualizações no conjunto de dados. Assim, as etapas apresentadas nesta seção – o tratamento, o relacionamento e a conversão dos dados – são de execução recorrente, podendo ser realizadas sempre que houver a necessidade de atualizar o conjunto de dados publicado como *Linked Data*. A periodicidade dessa atualização é definida pela instituição, assim como sua abrangência: tratamento, relacionamento e conversão de todos os registros que dão origem ao conjunto de dados ou somente daqueles modificados ou novos.

Destaca-se, ainda, que os procedimentos para o tratamento e o relacionamento dos dados podem ser realizados no próprio conjunto de dados já em RDF, após a conversão, o que depende das possibilidades das tecnologias utilizadas na disponibilização desses dados, conforme é abordado na seção seguinte.

5.4 Publicação dos dados

Após sua conversão para triplas em RDF, os dados de autoridade são publicados, tornando-se disponíveis para acesso e uso por qualquer um e para qualquer propósito. Para tanto, são realizadas:

- a descrição do conjunto de dados (Seção 5.4.1);
- a publicação do conjunto de dados em formatos acessíveis aos softwares (Seção 5.4.2);
- a publicação do conjunto de dados em formatos acessíveis aos humanos (Seção 5.4.3).

5.4.1 Descrição do conjunto de dados

A descrição de um conjunto de dados inclui informações sobre sua proveniência, por exemplo, seu título, sua cobertura, seus responsáveis e sua data de publicação ou atualização, sendo a disponibilização dessas informações uma das melhores práticas listadas pelo W3C: “A proveniência é um dos meios pelos quais os consumidores de um conjunto de dados julgam sua qualidade. Entender a origem e a história ajuda-os a determinar se confiam ou não nos dados e a prover um importante contexto de interpretação” (WORLD WIDE WEB CONSORTIUM, 2017, tradução nossa). O W3C descreve também nas melhores práticas a inclusão de informações sobre a licença, a qualidade, o propósito, a versão e o histórico de versões do conjunto de dados.

Uma vez que o conjunto de dados será utilizado por humanos e por softwares, sua descrição também deve estar em formatos acessíveis a estes e a aqueles. Existem vocabulários de classes e de propriedades para a descrição em RDF de conjuntos de dados, entre eles, o *Vocabulary of Interlinked Datasets (VOID)*,¹⁰⁹ o *Data Catalog Vocabulary (DCAT)*¹¹⁰ e o *Provenance Ontology (PROV-O)*¹¹¹. Algumas das classes e propriedades desses vocabulários estão exemplificadas na Figura 29.

¹⁰⁹ Disponível em: <<http://www.w3.org/TR/void/>>. Acesso em: 12 dez. 2017.

¹¹⁰ Disponível em: <<https://www.w3.org/TR/vocab-dcat/>>. Acesso em: 12 dez. 2017.

¹¹¹ Disponível em: <<https://www.w3.org/TR/prov-o/>>. Acesso em: 12 dez. 2017.

Figura 29 – Descrição em RDF de um conjunto de dados

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix void: <http://rdfs.org/ns/void#> .

<http://pessoas.com/dados>
  rdf:type dcat:Dataset ;
  dcterms:title "Arquivo de autoridade da Biblioteca X"@pt ;
  dcterms:creator <http://biblioteca-x.com/#biblioteca> ;
  dcterms:publisher <http://biblioteca-x.com/#biblioteca> ;
  dcterms:issued "2017-11-13"^^xsd:date ;
  dcterms:subject "Dados de autoridade"@pt , "Authority data"@en , "Biblioteca"@pt ;
  dcterms:description "Inclui dados extraídos de registros de autoridade representando pessoas."@pt ;
  dcterms:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
  void:exampleResource <http://pessoas.com/entidade/12345> ;
  void:dataDump <http://pessoas.com/dump/pessoas.rdf> ;
  void:dataDump <http://pessoas.com/dump/pessoas.json> ;
  void:dataDump <http://pessoas.com/dump/pessoas.ttl> ;
  void:sparqlEndpoint <http://pessoas.com/sparql> ;
  void:vocabulary <http://rdaregistry.info/Elements/a/> , <http://rdaregistry.info/Elements/z/> ,
    <http://www.w3.org/2004/02/skos/core#> .

```

Fonte: Elaborada pelo autor.

A descrição de um conjunto de dados o acompanha em suas diferentes formas de publicação em formatos acessíveis aos softwares e aos humanos, tratadas nas duas seções seguintes.

5.4.2 Dados acessíveis aos softwares

A publicação de conjuntos de dados em formatos acessíveis aos softwares ocorre com o uso de uma ou mais dessas estratégias: disponibilizar arquivos para *download* contendo o conjunto de dados em RDF, incluir as declarações em RDF nas páginas HTML, utilizar a negociação de conteúdo, utilizar uma API e utilizar um SPARQL *endpoint*. Essas estratégias são passíveis de uso simultâneo e com diferentes formatos de serialização das triplas (XML, N3, *N-Triples*, *Turtle*, JSON-LD, etc.).

A primeira estratégia consiste na disponibilização de arquivos contendo as triplas em RDF. Esses arquivos são os mesmos resultantes da conversão (Seção 5.3.3) ou, então, apresentam os dados serializados em outros formatos. Em razão da quantidade de triplas e para otimizar o uso e a atualização dos dados, a instituição pode optar por

dividi-los em vários arquivos, criando subconjuntos de dados, por exemplo, um subconjunto contendo somente os relacionamentos declarados com a propriedade *skos:exactMatch*, chamados conjuntos de *links (linksets)* (ALEXANDER et al., 2009, p. 3), ou, ainda, arquivos separados de acordo com as classes dos recursos: um para pessoas, um para entidades coletivas, outro para termos tópicos, etc. Os arquivos são disponibilizados em servidores da instituição ou em serviços como o *Datahub*, e o acesso a eles ocorre por meio de seu *download*. Esse modo de disponibilização apresenta vantagens para as instituições com poucos recursos tecnológicos, uma vez que, sozinho, não demanda tecnologias para a consulta aos dados em RDF, como uma API ou um SPARQL *endpoint*. Uma das desvantagens é que, para a consulta e o uso de uma declaração, é necessário o *download* do arquivo contendo todo o conjunto de dados.

A segunda estratégia consiste na inclusão de declarações em RDF nas páginas em HTML para que possam ser obtidas a partir do processamento dessas páginas. A inclusão de marcações em páginas HTML com o objetivo de explicitar certas partes de seu conteúdo que poderiam ser processadas automaticamente foi iniciada informalmente com os microformatos até o surgimento do RDFa, um formato de serialização do RDF que especificou como declarações poderiam ser incluídas em páginas HTML. A complexidade do RDFa, no entanto, levou ao surgimento do *Microdata*, com objetivo similar, porém, com a marcação mais simples (VAN HOOLAND; VERBORGH, 2015, p. 204). Entre as vantagens dessa estratégia estão o fornecimento de dados para os softwares e os humanos por meio de um único mecanismo (a página HTML) (STEPHENS, 2011, p. 246) e a melhor “interpretação” do conteúdo das páginas pelas ferramentas de busca (VAN HOOLAND; VERBORGH, 2015, p. 204).

A terceira estratégia consiste na utilização da negociação de conteúdo, abordada nas Seções 3.1 e 5.2.4, para prover as descrições em RDF serializadas em diferentes formatos ou páginas HTML.

As três estratégias até aqui apresentadas não dispõem, por si só, de interfaces para consulta aos dados, como ocorre com a quarta e quinta estratégias.

A quarta estratégia requer a utilização de uma *Application Programming Interface* (API), permitindo que outros softwares se comuniquem com os softwares utilizados na publicação dos dados, consultando e recuperando tais dados. Uma das desvantagens dessa estratégia é que cada API possui uma forma de consulta própria, de modo que seu utilizador precisa conhecer tal forma para conseguir utilizá-la.

A quinta e última estratégia compreende o uso de um SPARQL *endpoint*: uma interface que permite aos usuários, humanos ou softwares, consultar o conjunto de dados utilizando o protocolo SPARQL (SOUTHWICK, 2015, p. 9). A consulta a um SPARQL *endpoint* pode dar-se, aos usuários humanos, por meio de aplicações Web, enquanto que outros softwares podem consultá-lo a partir de APIs (YU, 2014, p. 267-268).

Essa interface, no entanto, depende do armazenamento dos dados em um banco de dados do tipo *triplestore*, que difere dos bancos de dados relacionais por permitir o armazenamento, de forma nativa, das declarações em RDF (SANTAREM SEGUNDO, 2014, p. 3877). Santarem Segundo (2014, p. 3876), levando em conta os ambientes *opensource* e as soluções livres, recomenda o uso do *Apache Jena*¹¹² para o início de estudos sobre a aplicação do SPARQL. O autor explica que

O Apache Jena é constituído de um conjunto de aplicações com destaque para algumas delas: suporte a elaboração de projetos usando arquivos em formato RDF; implementação de banco de dados persistente do tipo *triplestore* (bando de dados específico para formato RDF); ambientes de consulta SPARQL; mecanismos de inferência e suporte a ontologias, entre outras implementações diversas. (SANTAREM SEGUNDO, 2014, p. 3876, grifo do autor).

Outros softwares utilizados na criação de *triplestore* são: *Virtuoso*,¹¹³ *Blazegraph*¹¹⁴ e *Fuseki*¹¹⁵.

Embora o uso do SPARQL para consulta aos dados seja uma das melhores práticas para a publicação de *Linked Data* (BERNERS-LEE, 2006), os SPARQL *endpoints* nem sempre são mantidos pelos publicadores desses dados, como notado no Quadro 4, que sintetiza as iniciativas para a publicação de dados de autoridade apresentadas na Seção 4.2.

Uma das vantagens dos SPARQL *endpoints* em relação às APIs é o uso de uma linguagem padrão para a consulta, a SPARQL. Uma das desvantagens, no entanto, é que sua manutenção, no sentido amplo, ainda é custosa (VAN HOOLAND; VERBORGH, 2015, p. 220).

¹¹² Disponível em: <<https://jena.apache.org/>>. Acesso em: 12 dez. 2017.

¹¹³ Disponível em: <<https://virtuoso.openlinksw.com/>>. Acesso em: 12 dez. 2017.

¹¹⁴ Disponível em: <<https://www.blazegraph.com/>>. Acesso em: 12 dez. 2017.

¹¹⁵ Disponível em: <<https://jena.apache.org/documentation/fuseki2/>>. Acesso em: 12 dez. 2017.


5.4.3 Dados acessíveis aos humanos

As estratégias apresentadas na seção anterior permitem que softwares acessem os dados de autoridade publicados como *Linked Data* utilizando diferentes tecnologias. No entanto, o que todas essas estratégias têm em comum é a necessidade de se conhecer o conjunto de dados em questão, como aponta Santarem Segundo (2014, p. 3879, grifo do autor): “Para se recuperar informações de um determinado *dataset* é importante que se conheça a estrutura semântica utilizada por ele, os vocabulários associados e todo o contexto de que envolve a construção da informação que está relacionada”. Esses mesmos conjuntos de dados, acessíveis aos softwares, podem estar também acessíveis aos usuários humanos a partir de interfaces que não necessariamente demandam o conhecimento acerca da estrutura semântica empregada.

Conforme observado nas iniciativas descritas na Seção 4.2, a publicação de *Linked Data* de forma acessível aos humanos tem envolvido a criação de interfaces de consulta e de apresentação dos dados em páginas HTML. Essas interfaces se relacionam ao que Vila Suero e Gómez Pérez (2013, p. 598, tradução nossa) denominam exploração dos dados: “O objetivo da atividade de exploração é desenvolver aplicações e serviços que explorem os dados e forneçam interfaces ricas para os usuários finais e os desenvolvedores”.

Além da negociação de conteúdo e da inclusão de declarações em RDF nas páginas em HTML, que são estratégias para publicar dados acessíveis também aos humanos, os dados de autoridade podem ser apresentados também em agrupamentos criados a partir de características das entidades, por exemplo, uma página pode trazer tudo o que se relaciona a determinado ano ou local: pessoas que nasceram ou morreram naquele ano ou local, entidades coletivas atuantes naquele local, etc., como exemplificado na Figura 30 e na Figura 31, que trazem as páginas sobre o ano de 1973 e a cidade de Roma na iniciativa *data.bnf.fr*.

Figura 30 – Página sobre o ano de 1973



About data.bnf.fr | Contact

[<< The year 1972](#)
The period 1901-2000
The year 1974 >>

The year 1973

<ul style="list-style-type: none"> ▶ Related subjects (23) ▶ Authors born in 1973 (7 390) ▶ Died in 1973 (1 784) ▶ Organizations created in 1973 (1 423) ▶ End of activity in 1973 (563) 	<ul style="list-style-type: none"> ▶ Works created in 1973 (663) ▶ Work ended in 1973 (42) ▶ Performed in 1973 (951) ▶ Published in 1973 (52 569) ▶ Recorded in 1973 (327)
---	---

Related subjects 23 documents

- [Accords Kissinger-Lê Duc Tho \(1973\)](#)
Accords signés à Paris le 27-01-1973 par H. Kissinger et Lê Duc Tho à l'issue de la Conférence de Paris sur le Viet-Nam, prévoyant un cessez-le-feu, le retrait des troupes américaines, des élections libres au Viet-Nam du Sud, et l'ouverture d'une conférence de paix élargie, qui se tint à La Celle-Saint-Cloud (Yvelines) du 26-02 au 02-03-1973
- [Afghanistan -- Politique et gouvernement -- 1973-....](#)
- [Afghanistan -- 1973-....](#)
17-7-1973 : chute de la monarchie
- [Année 1973](#)

Fonte: Disponível em: <<http://data.bnf.fr/date/1973/>>.

Figura 31 – Página sobre a cidade de Roma

Rome (Italie)

Place : Roma (Italie) (italien)
 Geographic coordinates : E 12°30' / N 41°53'
 Geographic area : Latium (Italie, Europe)
 Related theme : Rome (Italie)

Related geographic records | Documents about this place (1 079 documents) | Other pages on Data.bnf.fr | Sources and references

Documents about this place

See all documents (1 079) | Digitized documents only (276)

- Cartes et plans (918)
- Livres (98)
- Images (55)
- Documents d'archives et manuscrits (catalogue général) (8)

Other pages in data.bnf.fr

- > Authors related to Rome (Italie)
- > Died in Rome (Italie)
- > Recorded in Rome (Italie)
- > Related journals in Rome (Italie)
- > Battle and treaty in Rome (Italie)
- > Born in Rome (Italie)
- > Organizations created in Rome (Italie)
- > Published in Rome (Italie)
- > Performed in Rome (Italie)

BnF Services

- Ask a librarian
- Come to the library
- Reproduction services

Resources from the BnF

- Search in
- Gallica
- Catalogue général
- BnF archives et manuscrits
- CNLJ - La Joie par les livres

External websites

- Search in

Fonte: Disponível em: <http://data.bnf.fr/15307086/rome_italie/>.

A escolha das melhores estratégias e tecnologias para a publicação de dados de autoridade acessíveis aos softwares e aos humanos depende da infraestrutura tecnológica da instituição, que inclui pessoal capacitado, hardware, software e conexão adequados.

5.5 Feedback e retroalimentação do arquivo de autoridade

A publicação de dados de autoridade como *Linked Data*, além de beneficiar a própria Web, beneficia, indiretamente, as práticas de catalogação e de controle de autoridade de uma instituição por meio do *feedback* que pode prover para essas práticas.

A modelagem dos dados e o mapeamento entre os metadados, o modelo de dados e as classes e propriedades em RDF atuam, secundariamente, como insumo para reflexões sobre as funções de cada um dos dados presentes nos registros de autoridade, sobre sua necessidade ou não no contexto da instituição e sobre a forma com que é registrado nos formatos utilizados. Durante o tratamento, o relacionamento e a conversão dos dados são identificadas inconsistências que podem ser utilizadas também para o aperfeiçoamento das práticas de catalogação e de controle de autoridade, desde a revisão das instruções utilizadas na instituição até a identificação da necessidade de capacitação da equipe responsável pelos dados de autoridade.

Concluída sua primeira publicação como *Linked Data*, os dados, antes restritos ao arquivo de autoridade na instituição, encontram-se prontos para o uso por qualquer um e para qualquer propósito na Web. No percurso entre o arquivo de autoridade da instituição e a Web, esses dados passaram por significativas melhorias, seja no tratamento, que objetivou a identificação e a correção de erros, seja no relacionamento com outros conjuntos de dados e no enriquecimento a partir desses relacionamentos. Essas melhorias fazem com que os dados de autoridade publicados como *Linked Data* sejam diferentes dos dados de autoridade presentes no arquivo da instituição.

Com o intuito de incorporar essas melhorias ao arquivo de autoridade da instituição, propõe-se, nesta etapa do modelo, que os dados publicados como *Linked Data* sejam convertidos de volta ao formato utilizado no arquivo de autoridade da instituição, de modo que essas melhorias possam constar também nesse arquivo.

Esse retorno dos dados ao arquivo de autoridade é necessário para que os usuários do sistema em que o arquivo está inserido, por exemplo, o sistema de gerenciamento de bibliotecas utilizado na instituição, possam se beneficiar das melhorias realizadas nos dados, tais como os relacionamentos e o enriquecimento. Esse retorno, no entanto, demanda o correto mapeamento das classes e propriedades em RDF para o padrão de metadados utilizado nos registros de autoridade, a codificação adequada de dados como os URIs¹¹⁶ e mecanismos de controle para evitar a perda de dados.

¹¹⁶ A codificação de URIs nos Formatos MARC 21 é assunto de discussões entre as instituições que mantêm esses formatos, como é observado no *Discussion Paper No. 2017-DP01: Use of \$0 and \$1 to Capture URIs*, disponível em <<https://www.loc.gov/marc/mac/2017/2017-dp01.html>>. Acesso em: 12 dez. 2017.

6 Considerações finais

Considerando o papel dos catálogos, com seus dados bibliográficos e de autoridade atuando para a localização, a identificação, a seleção, o acesso e a navegação nas coleções de recursos informacionais, e as possibilidades de uso das tecnologias da Web Semântica para a publicação de dados na Web, esta pesquisa foi iniciada com o questionamento “como publicar dados de autoridade como *Linked Data*?”. Ao longo desta tese, buscaram-se subsídios teóricos a partir da literatura nacional e internacional para a proposição de uma resposta a esse questionamento: um modelo para a publicação de dados de autoridade como *Linked Data*.

O percurso até essa proposição incluiu, inevitavelmente, a conceituação das funções dos dados de autoridade nos catálogos e seus benefícios nos ambientes de *Linked Data*, bem como das políticas, dos procedimentos e das tecnologias empregadas nesse ambiente. Para a condução desse percurso, foram definidos quatro objetivos específicos, cuja realização é ressaltada neste capítulo de considerações finais.

O primeiro objetivo específico, **conceituar as funções dos dados de autoridade nos catálogos bibliográficos**, foi alcançado no Capítulo 2. Nesse capítulo, ao discorrer sobre o desenvolvimento dos catálogos em livros, em fichas e digitais e sobre o arranjo de seus dados, buscou-se explicitar, em cada momento, as funções desempenhadas pelos dados de autoridade: permitir aos usuários encontrar, identificar, selecionar, contextualizar, justificar e explorar dados sobre entidades associadas a recursos informacionais. Considera-se aqui que essas funções, recentemente explicitadas para o ambiente digital pelos modelos conceituais FRAD e FRSAD, já se encontravam implícitas nos instrumentos de catalogação criados para o cenário analógico, como observado nas listas de autores do catalogador, descritas por Cutter, e nas fichas de autoridade incluídas na segunda edição preliminar do código da ALA.

O Capítulo 2 também exemplifica como as tecnologias utilizadas na criação dos catálogos analógicos tiveram impacto determinante na configuração desses catálogos, de suas possibilidades e limitações e das práticas de catalogação. A possibilidade de acrescentar novas entradas sem prejuízo à ordenação do catálogo, possibilidade essa oferecida pela tecnologia utilizada nos catálogos em fichas, favoreceu a proliferação dos pontos de acesso através dos quais os usuários poderiam encontrar determinado recurso informacional. Com as técnicas de reprodução de fichas, essa possibilidade

permitiu que as descrições completas dos recursos informacionais fossem replicadas e apresentadas em diversos locais dentro de um mesmo catálogo, favorecendo, assim, um acesso mais rápido a essas descrições por parte dos usuários. Além desses exemplos de possibilidades, têm-se também exemplos de limitações decorrentes das tecnologias utilizadas na criação dos catálogos. Nos catálogos em fichas, por exemplo, em que as descrições de cada edição ou versão de um recurso informacional eram apresentadas em unidades (fichas) individuais, as diferentes edições e versões de uma mesma obra poderiam estar dispersas ao longo do catálogo e não explicitamente relacionadas, como ocorria nos catálogos em livros.

Sobre a importância dos modelos conceituais FRAD e FRASAD, entende-se aqui que esses modelos, por definirem entidades, atributos, relacionamentos e tarefas dos usuários, apresentam uma sistematização do entendimento atual acerca do controle de autoridade nos catálogos. Essa sistematização, construída a partir de uma metodologia consolidada (a análise das entidades, atributos e relacionamentos) é de fundamental importância em tempos de mudanças como o vivenciado atualmente no domínio bibliográfico. Por meio dessa sistematização, assim como daquela apresentada no modelo FRBR, explicita-se que o domínio bibliográfico, com suas ferramentas de acesso à informação e seus instrumentos e processos de representação, fundamenta-se em teorias e não somente na atividade prática.

O segundo objetivo específico, **contextualizar as principais tecnologias da Web Semântica em relação à publicação de *Linked Data***, foi alcançado no Capítulo 3, em que foram abordados os URIs (utilizados na identificação de recursos), o modelo de dados RDF (utilizado como base para as declarações acerca dos recursos), o RDFS e a OWL (utilizados, no âmbito do *Linked Data*, para a descrição de vocabulários de classes e de propriedades e para o relacionamento de recursos).

O terceiro objetivo específico, **identificar os potenciais benefícios da publicação de dados de autoridade como *Linked Data* para os catálogos bibliográficos e para a Web**, foi alcançado na Seção 4.3, que sintetiza os principais benefícios em quatro segmentos: pesquisadores, estudantes e usuários; instituições; bibliotecários, arquivistas e curadores; e desenvolvedores e fornecedores de softwares; além dos benefícios gerais, potencialmente aplicáveis a toda Web. Com isso, a publicação como *Linked Data* é, acima de tudo, uma forma de redimensionamento dos dados de autoridade, de suas funções e escopos de atuação.

Considera-se aqui, que alguns desses benefícios são observados de forma concreta, como ocorre, por exemplo, com a melhoria dos processos internos de tratamento de dados descritos nas iniciativas *datos.bne.es* e *data.bnf.fr*. Outros, no entanto, permanecem difusos por diferentes motivos, entre eles (1) o estágio inicial das discussões acerca dos modelos de dados para a representação de dados bibliográficos e de autoridade no âmbito das práticas de *Linked Data*, (2) o estágio embrionário das iniciativas para a publicação desses dados, (3) o acesso às tecnologias da Web Semântica restrito principalmente às instituições de grande porte, tais como as bibliotecas nacionais, e (4) a falta de ferramentas que permitam aos “usuários finais” navegar pelos dados publicados como *Linked Data*.

O quarto objetivo específico, **identificar as principais iniciativas para a publicação de dados de autoridade como *Linked Data* e descrever os principais vocabulários que podem ser utilizados na publicação desses dados**, foi alcançado na Seção 4.2, que descreve de forma breve o *Library of Congress Linked Data Service*, o *datos.bne.es*, o *data.bnf.fr*, o VIAF e o AGROVOC, destacando suas políticas, seus procedimentos e suas tecnologias, e na Seção 4.1, que descreve o SKOS, o MADS/RDF e os *RDA Element Sets*. Sobre os vocabulários apresentados, considera-se aqui que a escolha por eles deve ocorrer principalmente com base no tipo de dado de autoridade a ser representado e nos requisitos de granularidade da instituição.

Sobre as iniciativas apresentadas, destaca-se a variedade de abordagens utilizadas para a modelagem dos dados, o que é reflexo das discussões iniciais acerca dos modelos de dados para a representação de dados bibliográficos e de autoridade no âmbito das práticas de *Linked Data*.

A partir do alcance desses quatro objetivos específicos, foi apresentado, no Capítulo 5, o **modelo para a publicação de dados de autoridade como *Linked Data*** com as etapas de planejamento; modelagem e mapeamento; tratamento, relacionamento e conversão; publicação; e *feedback* e retroalimentação, sobre as quais são apresentadas considerações finais.

As bibliotecas brasileiras e a literatura nacional da Ciência da Informação não estão discutindo as questões jurídicas relativas à publicação, ao acesso e ao reuso dos dados bibliográficos e de autoridade, mesmo fora do âmbito de sua publicação como *Linked Data*. Ainda que a Ciência da Informação, no Brasil, abarque questões relacionadas aos dados abertos, não foram localizados em sua literatura trabalhos sobre

a abertura dos dados bibliográficos ou de autoridade, sob a ótica da aplicação de licenças ou de outros dispositivos jurídicos. As bibliotecas brasileiras, principalmente as universitárias, mesmo aderindo às iniciativas de acesso aberto à informação científica e tecnológica, tais como os repositórios institucionais e as revistas de acesso aberto, não têm atribuído licenças aos dados presentes em seus próprios catálogos. Nesse sentido, o modelo, em sua etapa de planejamento, traz a oportunidade e a necessidade de discussão dessas questões.

A modelagem dos dados é o processo central da publicação de *Linked Data*, visto que as classes, as propriedades e os relacionamentos definidos durante essa modelagem possibilitarão – ou limitarão – os futuros usos e relacionamentos dos dados publicados. Nesse sentido, no que tange aos dados de autoridade, faz-se necessário que o modelo de dados utilizado assegure que, quando publicados como *Linked Data*, tais dados poderão desempenhar as funções que desempenham nos catálogos: permitir aos usuários encontrar, identificar, selecionar, contextualizar, justificar e explorar dados sobre entidades associadas a recursos informacionais.

A publicação de dados de autoridade como *Linked Data* requer o relacionamento com outros conjuntos de dados, seja a partir de uma entidade e um de seus atributos (por exemplo, uma pessoa e o seu local de nascimento), seja a partir de recursos que se referem ao mesmo objeto do mundo real (por exemplo, os registros de dois arquivos de autoridade que se referem à mesma pessoa). Esse segundo tipo de relacionamento, em determinados domínios, pode ser de difícil alcance devido à escassez de conjuntos de dados, o que se aplica ao cenário brasileiro atual no que tange aos dados de autoridade, visto que não foram identificados conjuntos desses dados publicados como *Linked Data* por instituições brasileiras. Com isso, um projeto dessa natureza que vier a surgir no Brasil estará, inicialmente, limitado a estabelecer relacionamentos com conjuntos de dados de outros países. Mas isso não impede que a instituição responsável pelo projeto antecipe-se, mantendo em seus registros de autoridade dados que futuramente possam ser utilizados para o relacionamento, por exemplo, o número de controle do registro na Biblioteca Nacional do Brasil.

A disponibilização dos dados e a criação de serviços a partir deles ainda são importantes obstáculos à publicação de *Linked Data*, uma vez que demandam não somente recursos humanos capacitados, mas uma infraestrutura tecnológica adequada para suportar os bancos de dados necessários e as consultas aos dados.

Há, ainda, a demanda por conhecimentos além dos tradicionalmente esperados dos bibliotecários, arquivistas e museólogos. Esses profissionais podem, por exemplo, selecionar conjuntos de dados, definir políticas, elaborar modelos conceituais, escolher vocabulários e estabelecer mapeamentos, mas precisam de profissionais da Ciência da Computação para concretizar o projeto, uma vez que os conhecimentos desses profissionais são requeridos principalmente para a conversão e a disponibilização dos dados em formatos acessíveis aos softwares e aos humanos. Assim, o modelo proposto pode auxiliar no diálogo entre os profissionais dessas áreas.

Além da discussão sobre tecnologias como as apresentadas nesta pesquisa e sua aplicação, a publicação de dados catalográficos, aqui não se falando especificamente de dados de autoridade, requer uma mudança da forma de pensamento dos responsáveis pela criação e pela gestão desses dados, mudança essa que se estende desde as decisões sobre a melhor forma de representação dos dados, por exemplo, qual formato será utilizado para a indicação das datas, até decisões políticas sobre o acesso e o uso dos dados por agentes externos, por exemplo, qual licença será aplicada ao conjunto de dados. São nessas decisões políticas que devem ser reforçados os papéis das instituições de patrimônio cultural, entre elas as bibliotecas, como importantes agentes proporcionadores da livre expressão, acesso e uso do conhecimento.

Com a aplicação do modelo proposto, os dados de autoridade podem ser publicados como *Linked Data* mesmo quando formatos como o MARC 21 para Dados de Autoridade continuam sendo utilizados na criação de registros. Ao mesmo tempo em que se nota essa possibilidade, questiona-se em que medida os dados de autoridade poderiam ser otimizados, enriquecidos e relacionados, e as práticas de controle de autoridade poderiam ser aperfeiçoadas se, nos sistemas utilizados para a criação dos registros, por exemplo, os sistemas de gerenciamento de bibliotecas, fossem utilizadas as tecnologias dos URIs, RDF, RDFS, OWL e SPARQL, no lugar do MARC 21 e das tecnologias conexas a ele.

Com esse questionamento, explicita-se o anseio de que o modelo proposto nesta pesquisa seja de necessidade temporária e transitória, diminuindo sua utilidade a partir do momento em que as melhores práticas para a publicação de *Linked Data* forem incorporadas aos sistemas utilizados na criação de registros de autoridade. Projetar ou prever como e quando se dará essa incorporação não está nos objetivos desta pesquisa,

uma vez que envolve não só questões tecnológicas, mas questões políticas e mercadológicas.

Também não esteve entre os objetivos desta pesquisa a aplicação do modelo em um conjunto de registros de autoridade, uma vez que a impossibilidade do alcance de um objetivo dessa natureza, devido à falta de recursos humanos e tecnológicos, comprometeria a exequibilidade da pesquisa dentro do cronograma e do prazo propostos.

O alcance do objetivo geral e dos objetivos específicos, ao longo desta pesquisa, conduziu a comprovação da tese de que o conhecimento acerca das funções dos dados de autoridade no contexto dos catálogos bibliográficos, aliado ao conhecimento das tecnologias da Web Semântica, possibilita a publicação dos dados de autoridade como *Linked Data* de modo que tais dados possam desempenhar, em nível macro na Web Semântica, as funções que desempenham em nível micro nos catálogos bibliográficos.

No Brasil, as publicações da Ciência da Informação têm contemplado questões iniciais relacionadas ao *Linked Data* e aos vocabulários de classes e de propriedades. Ao mesmo tempo, durante a realização desta pesquisa, não foram identificadas iniciativas brasileiras, no domínio bibliográfico, para a publicação de dados como *Linked Data*. Nesse sentido, essa pesquisa contribui não somente para o quadro nacional de publicações da Ciência da Informação, mas, principalmente, para os profissionais responsáveis pela criação e pela gestão de dados catalográficos, em especial, dos dados de autoridade.

Referências

ABELE, A. et al. **The Linking Open Data cloud diagram**. [S.l.: s.n.] Last updated: 2017-08-22. Disponível em: <<http://lod-cloud.net/>>. Acesso em: 12 dez. 2017.

AGENJO, X.; HERNÁNDEZ, F.; VIEDMA, A. Data aggregation and dissemination of authority records through Linked Open Data. In: WORLD LIBRARY AND INFORMATION CONGRESS, 77., 2011, San Juan. **Proceedings...** Den Haag: IFLA, 2011. Disponível em: <<https://www.ifla.org/past-wlic/2011/80-agenjo-en.pdf>>. Acesso em: 12 dez. 2017.

AGRICULTURAL INFORMATION MANAGEMENT STANDARDS. **AGROVOC multilingual agricultural thesaurus**. Roma, 2016a. Disponível em: <<http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>>. Acesso em: 23 ago. 2016.

AGRICULTURAL INFORMATION MANAGEMENT STANDARDS. **AGROVOC: about**. Roma, 2016b. Disponível em: <<http://aims.fao.org/standards/agrovoc/concept-scheme>>. Acesso em: 23 ago. 2016.

AGRICULTURAL INFORMATION MANAGEMENT STANDARDS. **AGROVOC: Linked Data**. Roma, 2015. Disponível em: <<http://aims.fao.org/standards/agrovoc/linked-open-data>>. Acesso em: 23 ago. 2016.

ALEXANDER, K. et al. Describing linked datasets: on the design and usage of void, the vocabulary of interlinked datasets. In: LINKED DATA ON THE WEB (LDOW2009), 2009, Madrid. **Proceedings...** Aachen: CEUR-WS, 2009. Disponível em: <http://events.linkeddata.org/ldow2009/papers/ldow2009_paper20.pdf>. Acesso em: 12 dez. 2017.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: <<http://hdl.handle.net/11449/103361>>. Acesso em: 12 dez. 2017.

ALVES, R. C. V.; SANTOS, P. L. V. A. da C. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013.

AMERICAN LIBRARY ASSOCIATION. **A.L.A. catalog rules: author and title entries**. Chicago: American Library Association, 1941. Disponível em: <<http://hdl.handle.net/2027/mdp.39015033890123>>. Acesso em: 12 dez. 2017.

ASSUMPCÃO, F. S.; SANTAREM SEGUNDO, J. E.; SANTOS, P. L. V. A. da C. RDA Element Sets and RDA Value Vocabularies: vocabularies for resource description in the Semantic Web. In: METADATA AND SEMANTICS RESEARCH (MTSR 2015), 9., 2015, Manchester. **Proceedings...** [S.l.]: Springer, 2015. p. 147-158. Disponível em: <http://dx.doi.org/10.1007/978-3-319-24129-6_13>. Acesso em: 12 dez. 2017.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. A importância do controle de autoridade: uma abordagem baseada nos objetivos e nas funções dos catálogos. In: ENCONTRO NACIONAL DE CATALOGADORES, 1., 2012, Rio de Janeiro. **Anais...** Rio de Janeiro: Fundação Biblioteca Nacional, 2012. Disponível em: <<http://hdl.handle.net/10760/18843>>. Acesso em: 12 dez. 2017.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. A utilização do Resource Description and Access (RDA) na criação de registros de autoridade para pessoas, famílias e entidades coletivas. **Encontros Bibli**, Florianópolis, v. 18, n. 37, p. 203-226, 2013a. Disponível em: <<http://dx.doi.org/10.5007/1518-2924.2013v18n37p203>>. Acesso em: 12 dez. 2017..

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. Conversão de registros em XML para MARC 21: um modelo baseado em XSLT. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 15., 2014, Belo Horizonte. **Anais...** Belo Horizonte: ECI/UFMG, 2014. p. 3802-3821. Disponível em: <<http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8>>. Acesso em: 12 dez. 2017.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. Linked Data no domínio bibliográfico: vocabulários para a publicação de dados de autoridade. In: ENCONTRO INTERNACIONAL DADOS, TECNOLOGIA E INFORMAÇÃO, 3., 2016, Marília. **Perspectivas e interdisciplinaridades em Ciência da Informação**. Marília: UNESP, 2016. p. 515-536. Disponível em: <<http://gpnti.marilia.unesp.br:8085/index.php/3DTI/3dti/announcement/view/18>>. Acesso em: 12 jan. 2016.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. Metadata Authority Description Schema (MADS): uma alternativa à utilização do formato MARC 21 para dados de autoridade. **Informação & Informação**, v. 18, n. 1, p. 106-126, 2013b. Disponível em: <<http://dx.doi.org/10.5433/1981-8920.2013v18n1p106>>. Acesso em: 12 dez. 2017.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C.; ZAFALON, Z. R. O controle de autoridade no domínio bibliográfico: os catálogos digitais. **Biblios**, v. 68, p. 21-33, 2017a. Disponível em: <<https://doi.org/10.5195/biblios.2018.342>>. Acesso em: 12 fev. 2018.

ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C.; ZAFALON, Z. R. O controle de autoridade no domínio bibliográfico: os catálogos em livros e em fichas. **Biblios**, n. 67, p. 84-98, 2017b. Disponível em: <<https://doi.org/10.5195/biblios.2017.341>>. Acesso em: 12 dez. 2017.

AULD, L. Authority control: an eighty-year review. **Library Resources & Technical Services**, v. 26, n. 4, p. 319-330, Oct./Dec. 1982. Disponível em: <<http://downloads.alcts.ala.org/lrts/lrtsv26no4.pdf>>. Acesso em: 30 nov. 2016.

AVRAM, H. D. **The MARC Pilot Project**: final report on a project sponsored by The Council on Library Resources, Inc. Washington, D.C.: Library of Congress, 1968. Disponível em: <<http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED029663>>. Acesso em: 30 nov. 2016.

BAKER, T.; COYLE, K.; PETIYA, S. Multi-entity models of resource description in the Semantic Web. **Library Hi Tech**, v. 32, n. 4, p. 562-582, 2014. Disponível em: <<http://dx.doi.org/10.1108/LHT-08-2014-0081>>. Acesso em: 12 dez. 2017.

BARBOSA, A. P. **Novos rumos da catalogação**. Rio de Janeiro: BNG/BRASILART, 1978.

BARITÉ, M. **Referenciales teóricos vigentes en el área de tratamiento temático de la información y su expresión metodológica**. Porto Alegre: ABEBD, 1998.

BERNERS-LEE, T. Is your Linked Open Data 5 star? In: BERNERS-LEE, T. **Linked Data**. Cambridge: W3C, 2010. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 30 nov. 2016.

BERNERS-LEE, T. **Linked Data**. Cambridge: W3C, 2006. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 24 set. 2013.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, May 2001, p. 29-37. Disponível em: <<http://www.scientificamerican.com/article/the-semantic-web/>>. Acesso em: 12 dez. 2017.

BETTENCOURT, A. M. **A representação da informação na Biblioteca Nacional: do documento tradicional ao digital**. Rio de Janeiro: Fundação Biblioteca Nacional, 2014. Disponível em: <<https://www.bn.br/producao/publicacoes/representacao-informacao-biblioteca-nacional-documento>>. Acesso em: 06 mar. 2016.

BIAGETTI, M. T. Catalogo alfabetico-classificato e catalogo dizionario: le considerazioni e le realizzazioni di Charles A. Cutter. In: NUOVO, A.; PETRUCCIANI, A.; RUFFINI, G. (Org.). **Pensare le biblioteche: studi e interventi offerti a Paolo Traniello**. Roma: Sinnos, 2008. p. 131-144.

BIBLIOTECA NACIONAL DE ESPAÑA. **Datos enlazados en la BNE**. Madri, 2016a. Disponível em: <<http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/index.html>>. Acesso em: 30 jul. 2016.

BIBLIOTECA NACIONAL DE ESPAÑA. **Datos.bne.es**: Guía de ayuda para la consulta. Madri, 2015. Disponível em: <<http://www.bne.es/webdocs/Inicio/Perfiles/Bibliotecarios/ayudadatosenlazados.pdf>>. Acesso em: 23 ago. 2016.

BIBLIOTECA NACIONAL DE ESPAÑA. **Modelo de datos**. Madri, 2016b. Disponível em: <<http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/Modelos/>>. Acesso em: 23 ago. 2016.

BIBLIOTHÈQUE NATIONALE DE FRANCE. **About data.bnf.fr**. Paris, 2016a. Disponível em: <<http://data.bnf.fr/about>>. Acesso em: 09 ago. 2016.

BIBLIOTHÈQUE NATIONALE DE FRANCE. **Semantic Web and data model**. Paris, 2016b. Disponível em: <<http://data.bnf.fr/about>>. Acesso em: 09 ago. 2016.

BOJĀRS, U.; ŽOGLA, A.; EGLĪTE, E. Linking library data for quality improvement and data enrichment. In: INTERNATIONAL CONFERENCE ON DATA MANAGEMENT TECHNOLOGIES AND APPLICATIONS (DATA-2015), 4., 2015, Colmar. **Proceedings...** Setúbal: SCITEPRESS, 2015. p. 184-188. Disponível em: <<http://dx.doi.org/10.5220/0005554901840188>>. Acesso em: 12 dez. 2017.

BORKO, H. Information Science: what is it? **American Documentation**, D.C., v. 19, n. 1, p. 3-5, 1968.

BOZZARELLI, O. **Authority control**: teorie, applicazioni e prospettive di sviluppo. 2004. 129 f. Tesi (Diploma in Teoria e tecniche della catalogazione e della classificazione) – Indirizzo Bibliotecari, Scuola speciale per archivisti e bibliotecari, Università degli Studi di Roma La Sapienza, Roma, 2004. Disponível em: <<http://hdl.handle.net/10760/18400>>. Acesso em: 12 dez. 2017.

BRANDT, S. Authority files for microcomputer databases. **Special Libraries**, v. 79, n. 4, p. 296-301, Fall 1988.

BRASIL. Lei nº 9610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. **Diário Oficial da União**, Brasília: Imprensa Nacional, 20 fev. 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9610.htm>. Acesso em: 11 dez. 2017.

BRITISH MUSEUM. **Catalogue of printed books in the British Museum**: volume I. London: Printed by order of the Trustees, 1841. Disponível em: <<http://hdl.handle.net/2027/umn.31951001998306w>>. Acesso em: 12 dez. 2017.

BROWN UNIVERSITY. **Catalogue of the library of Brown University in Providence, Rhode-Island**: with an index of subjects. Providence: Brown University, 1843. p. iii-xxi. Disponível em: <<https://archive.org/details/acataloguelibra12librgoog>>. Acesso em: 12 dez. 2017.

BURGER, R. H. **Authority work**: the creation, use, maintenance, and evaluation of authority records and files. Littleton: Libraries Unlimited, 1985.

BURT, P.; KINNUCAN, M. Information models and modeling techniques for information systems. **Annual Review of Information Science and Technology**, p. 175-208, 1990.

CALHOUN, K. A bird's eye view of authority control in cataloging. In: WORKSHOP ON THE COMPILATION, MAINTENANCE, AND DISSEMINATION OF TAXONOMIC AUTHORITY FILES (TAF) , 1998, Washington, DC. **Proceedings...** [S.l.: s.n.], 1998. Disponível em: <<http://researcharchive.calacademy.org/research/informatics/taf/proceedings/Calhoun.html>>. Acesso em: 25 fev. 2016.

CARACCIOLO, C. et al. The AGROVOC linked dataset. **Semantic Web**, v. 4, n. 3, p. 341-348, 2013. Disponível em: <<http://dx.doi.org/10.3233/SW-130106>>. Acesso em: 12 dez. 2017.

CARLSON, S.; SEELY, A. Using OpenRefine's reconciliation to validate local authority headings. **Cataloging & Classification Quarterly**, v. 55, n. 1, p. 1-11, 2017. Disponível em: <<https://doi.org/10.1080/01639374.2016.1245693>>. Acesso em: 12 dez. 2017.

CARPENTER, M. Main entry. In: SVENONIUS, E. (Org.). **The conceptual foundations of descriptive cataloging**. San Diego: Academic Press, 1989. p. 73-95.

CATARINO, M. E. Simple Knowledge Organization System: construindo sistemas de organização do conhecimento no contexto da Web Semântica. **Informação & Tecnologia**, João Pessoa, v. 1., n. 1. p. 17-28, jan./jun. 2014. Disponível em: <<http://periodicos.ufpb.br/ojs/index.php/itec/article/view/19307>>. Acesso em: 12 dez. 2017.

CATARINO, M. E.; SOUZA, T. B. A representação descritiva no contexto da web semântica. **Transinformação**, v. 24, n. 2, p. 78-90, 2012. Disponível em: <<http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/766>>. Acesso em: 12 dez. 2017.

CATMANDU: a data toolkit. [S.l.: s.n., 2017?]. Disponível: <<http://librecat.org/Catmandu/>>. Acesso em: 12 dez. 2017.

CHAN, L. M. **Cataloging and classification: an Introduction**. 2nd ed. New York: McGraw-Hill, 1994.

CHAN, L. M. **Cataloging and classification: an Introduction**. 3rd ed. Lanham: Scarecrow Press, 2007.

CHEN, Y-N. A RDF-based approach to metadata crosswalk for semantic interoperability at the data element level. **Library Hi Tech**, v. 33, n. 2, p. 175-194, 2015. Disponível em: <<http://dx.doi.org/10.1108/LHT-08-2014-0078>>. Acesso em: 12 dez. 2017.

CHEN, Y-N. A review of practices for transforming library legacy records into Linked Open Data. In: METADATA AND SEMANTICS RESEARCH (MTSR 2015), 9., 2015, Manchester. **Proceedings...** [S.l.]: Springer, 2017. p. 123-133. Disponível em: <https://doi.org/10.1007/978-3-319-70863-8_12>. Acesso em: 12 dez. 2017.

CINTRA, A. M. M. et al. **Para entender as linguagens documentárias**. São Paulo: Polis/APB, 1994.

CLACK, D. H. Authority control and linked bibliographic databases. **Cataloging & Classification Quarterly**, v. 8, n. 3/4, p. 35-46, 1988.

CLACK, D. H. **Authority control: principles, applications, and instructions**. Chicago: American Library Association, 1990.

CÓDIGO de catalogação anglo-americano. 2. ed., rev. 2002. São Paulo: FEBAB, 2004.

COLE, T. W. et al. Library MARC records into Linked Open Data: challenges and opportunities. **Journal of Library Metadata**, v. 13, n. 2-3, p. 163-196, 2013. Disponível em: <<http://dx.doi.org/10.1080/19386389.2013.826074>>. Acesso: 11 dez. 2017.

COLE, T. W. et al. Using linked open data to enhance the discoverability, functionality and impact of Emblematica Online. **Library Hi Tech**, v. 35, n. 1, p. 159-178, 2017. Disponível em: <<https://doi.org/10.1108/LHT-11-2016-0126>>. Acesso: 11 dez. 2017.

COYLE, K. **FRBR before and after**: a look at our bibliographic models. Chicago: American Library Association, 2016. Disponível em: <<http://kcoyle.net/beforeAndAfter/>>. Acesso em: 12 dez. 2017.

COYLE, K. Understanding the Semantic Web: bibliographic data and metadata. **ALA Library Technology Reports**, v. 46, n. 1, Jan. 2010. Disponível em: <<http://alatechsource.metapress.com/content/g212v1783607/>>. Acesso em: 01 jul. 2014.

CREATIVE COMMONS. **Sobre as Licenças**. [S.l., 2017?]. Disponível em: <<https://br.creativecommons.org/licencas/>>. Acesso em: 12 dez. 2017.

CUTTER, C. A. **Rules for a dictionary catalog**. 4th ed. rev. Washington, DC: Government Printing Office, 1904. Disponível em: <<https://archive.org/details/rulesforadictio06cuttgoog>>. Acesso em: 12 dez. 2017.

DEMO, P. **Metodologia científica em ciências sociais**. 3. ed. rev. e aum. São Paulo: Atlas, 1995.

DODDS, L.; DAVIS, I. **Linked data patterns**: a pattern catalogue for modelling, publishing, and consuming Linked Data. [S.l.], 2012. Disponível em: <<http://patterns.dataincubator.org/book/>>. Acesso em: 12 dez. 2017.

DODEBEI, V. L. D. **Tesouro**: linguagem de representação da memória documentária. Niterói: Intertexto, 2014.

DUNSIRE, G. et al. A reconsideration of mapping in a semantic world. In: DCMI INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2011, The Hague. **The Hague Proceedings**. [S.l.]: DCMI, 2011. Disponível em: <<http://dcpapers.dublincore.org/pubs/article/view/3622>>. Acesso em: 12 dez. 2017.

FERREIRA, J. A.; SANTOS, P. L. V. A. da C. O modelo de dados Resource Description Framework (RDF) e o seu papel na descrição de recursos. **Informação & Sociedade: estudos**, v. 23, n. 2, p. 13-23, 2013. Disponível em: <<http://dx.doi.org/10.6084/m9.figshare.1116375>>. Acesso em: 12 dez. 2017.

FIUZA, M. M. Funções e desenvolvimento do catálogo: uma visão retrospectiva. **Revista da Escola de Biblioteconomia da UFMG**, v. 9, n. 2, p. 139-158, set. 1980.

FRÍAS, J. A.; PÉREZ ALVAREZ, M. I. La ordenación de los asientos bibliográficos en los catálogos: hacia una normativa internacional. **Revista Española de Documentación Científica**, v. 18, n. 3, p. 283-297, 1995. Disponível em: <<http://dx.doi.org/10.3989/redc.1995.v18.i3.645>>. Acesso em: 12 dez. 2017.

FUSCO, E. **Aplicação dos FRBR na modelagem de catálogos bibliográficos digitais**. São Paulo: Cultura Acadêmica, 2011. Disponível em: <<http://hdl.handle.net/11449/109186>>. Acesso em: 12 dez. 2017.

GARRIDO ARILLA, M. R. **Teoría e historia de la catalogación de documentos**. Madrid: Síntesis, 1996.

GORMAN, M. Authority control in the context of bibliographic control in the electronic environment. **Cataloging & Classification Quarterly**, New York, v. 38, n. 3/4, p. 11-22, 2004. Disponível em: <http://dx.doi.org/10.1300/J104v38n03_03>. Acesso em: 12 dez. 2017.

GRACY, K. F.; ZENG, M. L.; SKIRVIN, L. Exploring methods to improve access to Music resources by aligning library Data with Linked Data: A report of methodologies and preliminary findings. **Journal of the American Society for Information Science and Technology**, v. 64, p. 2078-2099, 2013. Disponível em: <<http://dx.doi.org/10.1002/asi.22914>>. Acesso em: 12 dez. 2017.

GUENTHER, R. LC's Authorities and vocabularies web service: experimenting with Linked Data. AMERICAN LIBRARY ASSOCIATION MID-WINTER CONFERENCE, 2011, San Diego. **Papers & Presentations**. Washington, D.C.: Library of Congress, 2011. Disponível em: <<http://id.loc.gov/static/presentations/rgue-controlledvocabs-ala-mw2011-1.pdf>>. Acesso em: 12 dez. 2017.

GUERRINI, M. Verso una nuova struttura catalografica dei dati bibliografici. In: ROSSI, F.; TINTI, P. (Org.). **Belle contrade della memoria**: studi su documenti e libri in onore di Maria Gioia Tavoni. Bologna: Patron, 2009. Disponível em: <<http://hdl.handle.net/10760/13364>>. Acesso em: 12 dez. 2017.

GUIMARÃES, J. A. C. Abordagens teóricas de tratamento temático da informação (TTI): catalogação de assunto, indexação e análise documental. **Ibersid**: revista de sistemas de información y documentación, v. 3, p. 105-117, 2009. Disponível em: <<http://www.iversid.eu/ojs/index.php/iversid/article/view/3730>>. Acesso em: 12 dez. 2017.

HAGLER, R. **The bibliographic record and information technology**. 3rd ed. Chicago: American Library Association; Ottawa: Canadian Library Association, 1997.

HANSON, E. M. A beginner's guide to creating Library Linked Data: lessons from NCSU's Organization Name Linked Data Project. **Serials Review**, v. 40, n. 4, p. 251-258, 2014. Disponível em: <<http://dx.doi.org/10.1080/00987913.2014.975887>>. Acesso em: 12 dez. 2017.

HARPER, C. A.; TILLET, B. B. Library of Congress controlled vocabularies and their application to the Semantic Web. **Cataloging & Classification Quarterly**, v. 43, n. 3-4, p. 47-68, 2007. Disponível em: <https://doi.org/10.1300/J104v43n03_03>. Acesso em: 12 dez. 2017.

HEATH, T.; BIZER, C. **Linked Data**: evolving the web into a global data space. [S. l.]: Morgan & Claypool, 2011. Disponível em: <<http://linkeddatabook.com/editions/1.0/>>. Acesso em: 12 dez. 2017.

HERRERO PASCUAL, C. El control de autoridades. **Anales de Documentación**, n. 2, p. 121-136, 1999. Disponível em: <<http://revistas.um.es/analesdoc/article/view/2621>>. Acesso em: 12 dez. 2017.

HILLMANN, D. et al. RDA Vocabularies: process, outcome, use. **D-Lib Magazine**, v. 16, n. 1/2, Jan./Feb. 2010. Disponível em: <<http://www.dlib.org/dlib/january10/hillmann/01hillmann.html>>. Acesso em: 01 jul. 2014.

HITZLER, P.; KRÖTZSCH, M.; RUDOLPH, S. **Foundations of Semantic Web technologies**. Boca Raton: CRC Press, 2010.

IFLA STUDY GROUP ON THE FRBR. **Functional Requirements for Bibliographic Records**: final report. München: IFLA Section on Cataloguing, 1998. Disponível em: <<http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf>>. Acesso em: 12 dez. 2017.

IFLA WORKING GROUP ON FRANAR. **Functional Requirements for Authority Data**: a conceptual model. Den Haag: IFLA, 2013. Disponível em: <http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf>. Acesso em: 12 dez. 2017.

IFLA WORKING GROUP ON FRASAR. **Functional Requirements for Subject Authority Data (FRSAD)**: a conceptual model. Den Haag: IFLA, 2010. Disponível em: <<http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf>>. Acesso em: 12 dez. 2017.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. **Statement of international cataloguing principles**. Den Haag: IFLA, 2009. Disponível em: <http://www.ifla.org/files/cataloguing/icp/icp_2009-en.pdf>. Acesso em: 24 set. 2013.

JEWETT, C. C. Preface. In: BROWN UNIVERSITY. **Catalogue of the Library of Brown University in Providence, Rhode-Island**: with an index of subjects. Providence: Brown University, 1843. p. iii-xxi. Disponível em: <<https://archive.org/details/acataloguelibra12librgoog>>. Acesso em: 12 dez. 2017.

KORN, K.; OPPENHEIM, C. **Licensing Open Data: a practical guide**. [S.l.: s.n.], 2011. Disponível em: <http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf>. Acesso em: 12 dez. 2017.

LAMPERT, C. K.; SOUTHWICK, S. B. Leading to linking: introducing Linked Data to academic library digital collections. **Journal of Library Metadata**, v. 13, n. 2-3, p. 230-253, 2013. Disponível em: <<http://dx.doi.org/10.1080/19386389.2013.826095>>. Acesso em: 12 dez. 2017.

LIBRARY OF CONGRESS. **Authorities, a MARC format**: specifications for magnetic tapes containing authority records. Preliminar edition. Washington: Library of Congress, 1976. Disponível em: <<http://hdl.handle.net/2027/uiug.30112018626959>>. Acesso em: 12 dez. 2017.

LIBRARY OF CONGRESS. **Library of Congress Linked Data Service**. Washington, D.C., 2016a. Disponível em: <<http://id.loc.gov/>>. Acesso em: 30 ago. 2016.

LIBRARY OF CONGRESS. **Library of Congress Linked Data Service**: about. Washington, D.C., 2016b. Disponível em: <<http://id.loc.gov/about/>>. Acesso em: 30 ago. 2016.

LIBRARY OF CONGRESS. **Library of Congress Linked Data Service**: technical center: searching/querying. Washington, D.C., 2016c. Disponível em: <<http://id.loc.gov/techcenter/searching.html>>. Acesso em: 30 ago. 2016.

LIBRARY OF CONGRESS. **MADS/RDF Primer**. Washington, D.C., 2015. Disponível em: <<http://www.loc.gov/standards/mads/rdf/20151028/>>. Acesso em: 24 maio 2016.

LUBETZKY, S. Crisis in the catalog. In: SVENONIUS, E.; MCGARRY, D. (Eds.). **Seymour Lubetzky**: writings on the classical art of cataloging. Englewood: Libraries Unlimited, 2001. Paper read before the Large Libraries Round Table of the Catalog Section, ALA, San Francisco, June 20, 1939. Originalmente publicado em: *Cataloguers' and Classifiers' Yearbook* 8, p. 48-54, 1940.

LUBETZKY, S. The vicissitudes of ideology and technology in anglo-american cataloging since Panizzi and a prospective reformation of the catalog for the next century. In: SVENONIUS, E.; MCGARRY, D. (Eds.). **Seymour Lubetzky**: writings on the classical art of cataloging. Englewood: Libraries Unlimited, 2001. Originalmente publicado em: *The future of cataloging: insights from the Lubetzky Symposium*, April 18, 1998. Chicago: American Library Association, 2000.

MALINCONICO, M. S. The role of a machine based authority file in an automated bibliographic system. In: CARPENTER, M.; SVENONIUS, E. (Ed.). **Foundations of cataloging**: a sourcebook. Littleton: Libraries Unlimited, 1985. Originalmente publicado em: *Automation in Libraries: Papers Presented at the CACUL Workshop on Library Automation*, Winnipeg, June 22-23, 1974. Ottawa: Canadian Library Association, 1975.

MALMSTEN, M. Exposing Library Data as Linked Data. In: IFLA satellite preconference sponsored by the Information Technology Section Emerging trends in technology: Libraries between Web 2.0, the Semantic Web and search technology, 2009, Florence. **Proceedings...** Den Haag: IFLA, 2009. Disponível em: <http://wtlab.um.ac.ir/images/e-library/linked_data/other/Exposing%20Library%20Data%20as%20Linked%20Data.pdf>. Acesso em: 12 dez. 2017.

MARAIS, H. **Authority control in an academic library consortium using a union catalogue maintained by a central office for authority control**. 2004. 310 f. Tese (Doctor of literature and philosophy in the subject Information Science) - University of South Africa, Petroria, 2004. Disponível em: <<http://hdl.handle.net/10500/2546>>. Acesso em: 12 dez. 2017.

MARCONDES, C. H. "Linked data" – dados interligados - e interoperabilidade entre arquivos, bibliotecas e museus na web. **Encontros Bibli**, v. 17, n. 34, p. 171-192, 2012. Disponível em: <<http://dx.doi.org/10.5007/1518-2924.2012v17n34p171>>. Acesso em: 12 dez. 2017.

MARCONDES, C. H. Tecnologias LOD e a publicação e interligação de acervos digitais de arquivos, bibliotecas e museus na Web. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 27., 2017, Fortaleza. **Anais...** São Paulo: FEBAB, 2017. Disponível em: <<https://portal.febab.org.br/anais/article/view/1930>>. Acesso em: 12 dez. 2017.

MAXWELL, R. L. **Maxwell's guide to authority work**. Chicago: American Library Association, 2002.

MCCALLUM, S. H. MARC: keystone for library automation. **IEEE Annals of the History of Computing**, v. 24, n. 2, p. 34-49, Apr./Jun. 2002. Disponível em: <<http://dx.doi.org/10.1109/MAHC.2002.1010068>>. Acesso em: 12 dez. 2017.

MEY, E. S. A.; SILVEIRA, N. C. **Catálogo no plural**. Brasília: Briquet de Lemos/Livros, 2009.

MORSLED, A. et al. Thesaurus Alignment for Linked Data Publishing. In: DCMI INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2011, The Hague. **The Hague Proceedings**. [S.l.]: DCMI, 2011. p. 37-46. Disponível em: <<http://dcpapers.dublincore.org/pubs/article/view/3623>>. Acesso em: 12 dez. 2017.

NOWACK, B. The Semantic Web: not a piece of cake... In: NOWACK, B. **Bnode**. [S.l.: s.n.], 2009. Disponível em: <<http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake>>. Acesso em: 01 dez. 2016.

OBILLE, K. L. B. AACR2 to RDA: a paradigm shift? **Journal of Philippine Librarianship**, v. 32, p. 1-12, 2012. Disponível em: <<http://journals.upd.edu.ph/index.php/jpl/article/view/3616>>. Acesso em: 01 dez. 2016.

ONLINE COMPUTER LIBRARY CENTER. **History of VIAF**. Dublin, Ohio, 2016a. Disponível em: <<http://www.oclc.org/viaf/history.en.html>>. Acesso em: 31 ago. 2016.

ONLINE COMPUTER LIBRARY CENTER. **VIAF: the Virtual International Authority File**. Dublin, Ohio, 2016b. Disponível em: <<http://viaf.org/>>. Acesso em: 31 ago. 2016.

ONLINE COMPUTER LIBRARY CENTER. **VIAF**: the Virtual International Authority File. Dublin, Ohio, 2014. Disponível em: <<http://www.oclc.org/viaf.en.html>>. Acesso em: 31 ago. 2016.

OPEN KNOWLEDGE FOUNDATION. **Principles on Open Bibliographic Data**. London, 2011.

OPEN KNOWLEDGE INTERNATIONAL. **ODC Public Domain Dedication and Licence (PDDL)**. London, [2008?]. Disponível em: <<https://opendatacommons.org/licenses/pddl/1.0/>>. Acesso em: 12 dez. 2017.

OPEN KNOWLEDGE INTERNATIONAL. **Open Data Commons Attribution License (ODC-By) v1.0**. London, [2010?]. Disponível em: <<https://opendatacommons.org/licenses/by/1.0/>>. Acesso em: 12 dez. 2017.

OPEN KNOWLEDGE INTERNATIONAL. **Open Database License (ODbL) v1.0**. London, [2014?]. Disponível em: <<https://opendatacommons.org/licenses/odbl/1.0/>>. Acesso em: 12 dez. 2017.

OPEN KNOWLEDGE INTERNATIONAL. **Open Definition 2.1**. London: [2015?]. Disponível em: <<http://opendefinition.org/od/2.1/en/>>. Acesso em: 12 dez. 2017.

ORTEGA, C. D. Do princípio monográfico à unidade documentária: exploração dos fundamentos da Catalogação. **Liinc em Revista**, v. 7, n. 1, p. 43-60, mar. 2011. Disponível em: <<http://liinc.revista.ibict.br/index.php/liinc/article/view/402>>. Acesso em: 01 dez. 2016.

ORTEGA, C. D.; LARA, M. L. G. de. A noção de estrutura e os registros de informação dos sistemas documentários. **Transinformação**, v. 227, n. 1, p. 7-17, jan./abr. 2010. Disponível em: <<http://dx.doi.org/10.1590/S0103-37862010000100001>>. Acesso em: 12 dez. 2017.

PANIZZI, A. Rules for the compilation of the catalogue. In: BRITISH MUSEUM. **Catalogue of printed books in the British Museum**: volume I. London: J. B. Nichols and Son Printers, 1841. p. v-ix. Disponível em: <<http://hdl.handle.net/2027/umn.31951001998306w>>. Acesso em: 12 dez. 2017.

PASTOR SÁNCHEZ, J. A.; MARTÍNEZ MÉNDEZ, F. J.; RODRÍGUEZ MUÑOZ, J. V. Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. **El profesional de la información**, v. 21, n. 3, p. 245-253, mayo/junio 2012. Disponível em: <<http://dx.doi.org/10.3145/epi.2012.may.04>>. Acesso em: 12 dez. 2017.

PATTUELLI, M. C. Personal name vocabularies as Linked Open Data: a case study of jazz artist names. **Journal of Information Science**, v. 38, n. 6, p. 558-565, 2012. Disponível em: <<http://dx.doi.org/10.1177/0165551512455989>>. Acesso em: 12 dez. 2017.

RAMALHO, R. A. S. **Desenvolvimento e utilização de ontologias em bibliotecas digitais: uma proposta de aplicação**. 2010. 145 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: <<http://hdl.handle.net/11449/103376>>. Acesso em: 12 dez. 2017.

RAMALHO, R. A. S. O modelo de dados SKOS: novas perspectivas no âmbito da representação do conhecimento. In: Encontro Nacional de Pesquisa em Ciência da Informação, 17., 2016, Salvador. **Anais...** Salvador: PPGCI/UFBA, 2016. Disponível em: <<http://www.ufpb.br/evento/lti/ocs/index.php/enancib2016/enancib2016/paper/view/4078>>. Acesso em: 12 dez. 2017.

RAYWARD, W. B. A history of computer applications in libraries: prolegomena. **IEEE Annals of the History of Computing**, v. 24, n. 2, p. 4-15, Apr./Jun. 2002. Disponível em: <<http://dx.doi.org/10.1109/MAHC.2002.1010066>>. Acesso em: 06 mar. 2016.

RDA STEERING COMMITTEE. **Resource Description and Access**. Chicago: ALA Publishing, 2016. Disponível em: <<http://www.rdatoolkit.org/>>. Acesso em: 17 ago. 2016.

RODRÍGUEZ DONCEL, V.; GÓMEZ PÉREZ, A.; MIHINDUKULASOORIYA, N. Rights declaration in linked data. In: INTERNATIONAL CONFERENCE ON CONSUMING LINKED DATA, 4., 2013, Sydney. **Proceedings...** Aachen: CEUR Workshop Proceedings, 2013. Disponível em: <http://ceur-ws.org/Vol-1034/RodriguezDoncelEtAl_COLD2013.pdf>. Acesso em: 12 dez. 2017.

SANTAREM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 16., 2015, João Pessoa. **Anais...** João Pessoa: ANCIB, 2015. Disponível em: <<http://inseer.ibict.br/ancib/index.php/tpbci/article/viewArticle/207>>. Acesso em: 06 mar. 2016.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a recuperação de dados usando SPARQL. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 15., 2014, Belo Horizonte. **Anais...** Belo Horizonte: ECI/UFMG, 2014. p. 3863-3882. Disponível em: <<http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8>>. Acesso em: 12 dez. 2017.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Web Semântica e Ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação e Informação**, v. 21, n. 2, p. 217-244, 2016. Disponível em: <<http://dx.doi.org/10.5433/1981-8920.2016v21n2p217>>. Acesso em: 12 dez. 2017.

SANTOS, P. L. V. A. da C.; ALVES, R. C. V. Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0. **DataGramZero**: Revista de Ciência da Informação, v. 10, n. 6, dez. 2009. Disponível em: <http://www.dgz.org.br/dez09/Art_04.htm>. Acesso em: 06 mar. 2016.

SANTOS, P. L. V. A. da C.; PEREIRA, A. M. **Catálogo**: breve histórico e contemporaneidade. Niterói: Intertexto, 2014.

SANTOS, P. L. V. A. da C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação**, v. 42, n. 2, p. 199-209, maio/ago. 2013. Disponível em: <<http://revista.ibict.br/ciinf/article/view/1382>>. Acesso em: 12 dez. 2017.

SANTOS, R.; MANCHADO, A.; VILA SUERO, D. Datos.bne.es: a LOD service and a FRBR-modelled access into the library collections. In: WORLD LIBRARY AND INFORMATION CONGRESS, 2015, Cape Town. **Proceedings...** Den Haag: IFLA, 2015. Disponível em: <<http://library.ifla.org/id/eprint/1085>>. Acesso em: 12 dez. 2017.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, v. 24, n. 1, 1995. Disponível em: <<http://revista.ibict.br/ciinf/article/view/608>>. Acesso em: 12 dez. 2017.

SAYÃO, L. F. Modelos teóricos em ciência da informação: abstração e método científico. **Ciência da Informação**, v. 30, n. 1, p. 82-91, jan./abr. 2001. Disponível em: <<http://dx.doi.org/10.1590/S0100-19652001000100010>>. Acesso em: 12 dez. 2017.

SAYERS, W. C. B.; STEWART, J. D. **The Card catalogue**: a practical manual for public and private libraries: including other uses of cards in libraries. London: Grafton & Company, 1913. Disponível em: <<https://archive.org/details/cardcataloguepra00sayerich>>. Acesso em: 12 dez. 2017.

SERRA, L. G.; SANTAREM SEGUNDO, J. E. O catálogo da biblioteca e o linked data. **Em Questão**, Porto Alegre, v. 23, n. 2, p. 167-185, maio/ago. 2017. Disponível em: <<http://dx.doi.org/10.19132/1808-5245232.167-185>>. Acesso em: 12 dez. 2017.

SIMON, A. et al. Publishing bibliographic records on the web of data: opportunities for the BnF (French National Library). In: CIMIANO, P. et al. (Ed.). **The Semantic Web: Semantics and Big Data**. Berlin: Springer, 2013. Disponível em: <http://dx.doi.org/10.1007/978-3-642-38288-8_38>. Acesso em: 12 dez. 2017.

SOUSA, B. P. de; SALDANHA, G. S.; TOLENTINO, V. de S. Possibilidades reflexivas sobre gênero na estruturação do ponto de acesso na Catalogação: entre a delimitação das regras e a amplitude da representação. In: INTERNATIONAL ISKO CONFERENCE, 14., 2017, Rio de Janeiro. **Memória, tecnologia e cultura na organização do conhecimento**. Recife: Ed. UFPE, 2017. Disponível em: <<http://isko-brasil.org.br/wp-content/uploads/2013/02/livro-ISKO-2017.pdf>> Acesso em: 12 dez. 2017.

SOUTHWICK, S. B. A guide for transforming digital collections metadata into Linked Data using open source technologies. **Journal of Library Metadata**, v. 15, n. 1, p. 1-35, 2015. Disponível em: <<http://dx.doi.org/10.1080/19386389.2015.1007009>>. Acesso em: 12 dez. 2017.

STEPHENS, O. Mashups and open data in libraries. *Serials*, v. 24, n. 3, p. 245-250, 2011. Disponível em: <<http://doi.org/10.1629/24245>>. Acesso em: 12 dez. 2017.

STROMGREN, Pip. **Charles Ammi Cutter**: library systematizer extraordinaire. Northampton: Forbes Library, 2004. Disponível em: <<http://forbeslibrary.org/info/library-history/charles-ammi-cutter/>>. Acesso em: 12 dez. 2017.

SUMMERS, E. et al. LCSH, SKOS and Linked Data. In: DCMI INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2008, Berlin. **Proceedings...** [S.l.]: DCMI, 2011. Disponível em: <<http://dcpapers.dublincore.org/pubs/article/viewFile/916/912>>. Acesso em: 12 dez. 2017.

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge: MIT Press, 2000.

TAIT, J. A. **Authors and titles**: an analytical study of the author concept in codes of cataloguing rules in the English language, from that of the British Museum in 1841 to the Anglo-American Cataloguing Rules 1967. London: Bingley, 1969.

TAYLOR, A. G. Authority files in online catalogs: an investigation of their value. **Cataloging & Classification Quarterly**, v. 4, n. 3, p. 1-17, 1984.

TAYLOR, A. G. **Introduction to cataloging and classification**. 10th ed. Westport: Libraries Unlimited, 2006.

TAYLOR, A. G.; JOUDREY, D. N. **The organization of information**. 3rd ed. Westport: Libraries Unlimited, 2009.

THARANI, K. Linked Data in libraries: a case study of harvesting and sharing bibliographic metadata with BIBFRAME. **Information Technology and Libraries**, v. 34, n. 1, p. 5-19, 2015. Disponível: <<https://doi.org/10.6017/ital.v34i1.5664>>. Acesso em: 12 dez. 2017.

TILLET, B. B. A summary of the treatment of bibliographic relationships in cataloging rules. **Library Resources and Technical Services**, v. 35, n. 4, p. 393-405, Oct. 1991. Disponível em: <<http://downloads.alcts.ala.org/lrts/lrtsv35no4.pdf>>. Acesso em: 12 dez. 2017.

TILLET, B. B. Bibliographic structures: the evolution of catalog entries, references, and tracings. In: SVENONIUS, Elaine (Org.). **Conceptual foundations of descriptive cataloging**. San Diego: Academic Press, 1989, p. 149-165.

TILLET, B. B. Prólogo. In: JIMÉNEZ PELAYO, J.; GARCÍA BLANCO, R. **El catálogo de autoridades**: creación y gestión en unidades documentales. Gijón: Trea, 2002. p. 7-10.

TILLET, B. B. The history of linking devices. **Library Resources and Technical Services**, v. 36, n. 1, p. 23-36, Jan. 1992. Disponível em: <<http://downloads.alcts.ala.org/lrts/lrtsv36no1.pdf>>. Acesso em: 12 dez. 2017.

VAN DER GRAAF, M.; WAAIJERS, L. **Authority files**: Breaking out of the library silo to become signposts for research information. Bristol: Knowledge Exchange, 2014. Disponível em: <<http://knowledge-exchange.info/event/authority-files>>. Acesso em: 12 dez. 2017.

VAN HOOLAND, S.; VERBORGH, R. **Linked Data for libraries, archives and museums**: how to clean, link and publish your metadata. Londres: Facet, 2015.

VILA SUERO, D.; GÓMEZ PÉREZ, A. datos.bne.es and MARiMbA: an insight into Library Linked Data. **Library Hi Tech**, v. 31, n. 4, p. 575-601, 2013. Disponível em: <<http://dx.doi.org/10.1108/LHT-03-2013-0031>>. Acesso em: 12 dez. 2017.

VILA SUERO, D.; VILLAZÓN TERRAZAS, B.; GÓMEZ PÉREZ, A. datos.bne.es: a Library Linked Data dataset. **Semantic Web**, p. 1-7, 2012. Disponível em: <<http://www.semantic-web-journal.net/content/datosbnees-library-linked-data-dataset>>. Acesso em: 12 dez. 2017.

W3C LIBRARY LINKED DATA INCUBATOR GROUP. **Final report**. Cambridge, 2011. Disponível em: <<https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>>. Acesso em: 12 dez. 2017.

W3C LIBRARY LINKED DATA INCUBATOR GROUP. **W3C Library Linked Data Incubator Group**. Cambridge, 2012. Disponível em: <<http://www.w3.org/2005/Incubator/lld/>>. Acesso em: 12 dez. 2017.

WILLER, M.; DUNSIRE, G. **Bibliographic information organization in the semantic web**. Oxford: Chandos Publishing, 2013.

WORLD WIDE WEB CONSORTIUM. **Cool URIs for the Semantic Web**. Cambridge, 2008. Disponível em: <<http://www.w3.org/TR/2008/NOTE-cooluris-20081203/>>. Acesso em: 12 dez. 2017.

WORLD WIDE WEB CONSORTIUM. **Data on the Web Best Practices**. Cambridge, 2017. Disponível em: <<https://www.w3.org/TR/2017/REC-dwbp-20170131/>>. Acesso em: 12 dez. 2017.

WORLD WIDE WEB CONSORTIUM. **OWL 2 Web Ontology Language**: document overview. 2nd ed. Cambridge, 2012a. Disponível em: <<https://www.w3.org/TR/2012/REC-owl2-overview-20121211/>>. Acesso em: 18 maio 2016.

WORLD WIDE WEB CONSORTIUM. **OWL 2 Web Ontology Language**: quick reference guide. 2nd ed. Disponível. Cambridge, 2012b. em: <<https://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/>>. Acesso em: 18 maio 2016.

WORLD WIDE WEB CONSORTIUM. **RDF Primer**. Cambridge, 2014a. Disponível em: <<https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>>. Acesso em: 07 mar. 2016.

WORLD WIDE WEB CONSORTIUM. **RDF Schema 1.1**. Cambridge, 2014b. Disponível em: <<https://www.w3.org/TR/2014/REC-rdf-schema-20140225/>>. Acesso em: 18 maio 2016.

WORLD WIDE WEB CONSORTIUM. **SKOS Simple Knowledge Organization System: primer**. Cambridge, 2009. Disponível em: <<http://www.w3.org/TR/skos-primer/>>. Acesso em: 01 set. 2015.

WORLD WIDE WEB CONSORTIUM. **Vocabularies**. Cambridge, 2015. Disponível em: <<http://www.w3.org/standards/semanticweb/ontology>>. Acesso em: 18 maio 2016.

YU, L. **A developer's guide to the Semantic Web**. 2nd ed. Heidelberg: Springer, 2014.

ZAPOUNIDOU, S.; SFAKAKIS, M.; PAPTAEODOROU, C. Representing and integrating bibliographic information into the Semantic Web: A comparison of four conceptual models. **Journal of Information Science**, v. 43, n. 4, p. 525-553, 2017. Disponível em: <<https://doi.org/10.1177/0165551516650410>>. Acesso em: 12 dez. 2017.