Proposal of a Metadata Application Profile for Technical Reports

Morgana Andrade[1(✉)] and Ana Alice Baptista[2]

[1] Espirito Santo Federal University, Vitória, Espírito Santo, Brazil
morganaandrade@hotmail.com
[2] University of Minho, Guimarães, Portugal

Abstract. Technical reports may be elaborated to serve multiple objectives and they constitute important sources of information for researchers and organizations. However their discovery and retrieval may be compromised due to particular characteristics related to the way they are described. A metadata application profile is recognized for providing increased interoperability and semantics between information systems, favouring discovery. In this article we report on the process of developing an application profile for technical reports and present the respective profile. This application profile includes elements from metadata vocabularies already consolidated in the context of scholarly publication and also elements from a specific metadata vocabulary developed within the scope of this work, *techrap*. The final product of this study are the constraints matrices with 7 classes and 89 properties. The results of this work are useful for metadata librarians and managers of online digital respositories.

Keywords: Technical report · Metadata application profile · Linked data

1 Introduction

A technical report may be defined as a "[…] a document that formally reports on the results or progress of research and development or that describes the state of a technical or scientific issue" [1]. It may be classified as: "technical-scientific, of travel and participation in events, of traineeship, of technical visit, administrative, for special and progressive purposes" and is drawn up for the function or under the responsibility of an organization or person [2, 3]. The first records of publication and use of reports as a means of scholarly communication date back to the period of 1909–1915, by the National Aeronautics and Space Administration - NASA [3].

Technical reports, identified by some authors as grey literature, may contribute to the development of teaching and research activities [4] but, due to their characteristics, they require specific treatment for their management and retrieval [5]. According to González de Gomez and Machado [6], access to this type of resources may be hampered in part by: the independence with which they are developed, i.e., the research activities described in a report may have been developed within an institution to which the authors are not bound; confidentiality issues; intellectual property rights; dissemination and reproduction restrictions [3, 7].

E. Garoufallou et al. (Eds.): MTSR 2019, CCIS 1057, pp. 175–186, 2019.
https://doi.org/10.1007/978-3-030-36599-8_15

Currently, technical reports and other types of grey literature are being stored using a variery of information systems, mainly in scholarly digital repositories (SDR). Of the 3,519 SDR in the world, 35% store this type of publication [8].

González de Gomez and Machado [6] have shown concern about how these documents are described and how this influences their organization and retrieval. They have proposed a set of metadata elements for describing reports based on MARC 21. On the other hand, Zou et al. [9] suggest elements for metadata schemas based on Functional Requirements for Bibliographic Records (FRBR). Turner, Liddy, Bradley, and Wheatley [10] developed a model for locating grey literature documents relating to public health interventions. They presented metadata elements that describe the organization, intervention type, methods (date/duration, setting and target population), outcomes and documents (document type and bibliographic elements).

Taking as a basis relevant literature on metadata and Linked Data, especially the one related to technical reports, in this paper we propose a Metadata Application Profile (MAP) for technical reports in the context of SDR. We consider both the physical and thematic representation. The creation and availability of MAP contributes to the uniformization and interrelation of resource descriptions, enhancing the semantic interoperability between different information systems. The results of this work are useful for metadata librarians and SDR managers.

This paper is structured as follows: Sect. 2 provides an overview of MAPs and Resource Description Framework (RDF); Sect. 3 presents the methodological procedures for the development of the MAP. The MAP for technical reports (TechRAP), is presented in Sect. 4. Section 5 presents the final considerations.

2 Metadata Application Profiles

The large amount of existing metadata schemas and vocabularies has created the conditions for the emergence of developments aimed at their creation, relation, recording, access, use and reuse. Among these are the MAP [11].

In this article, the term “application profile” is used such as defined by Heery and Patel [12]: “[…] schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application”. Although in the present case we are not thinking about a “particular local application”, it seems to us that this is still the most interesting definition of MAP as it focuses on the final scheme and does not include intermediate documents resulting from other phases of the MAP development process as, for example, the Singapore Framework does.

From the literature it is possible to draw the history of the concept of application profile. It is clear that the article by Heery and Patel [12] was a milestone for the definition of the concept currently adopted by the DCMI community (http://repositorio.ufes.br/handle/10/11414). In Semantic Web applications, the use of a MAP is recommended at least when it is necessary to use properties drawn from different metadata vocabularies, including own vocabularies [12, 13], as long as these vocabularies are defined in RDF and follow the generic model for metadata records (DCAM) [12, 14].

For clarification, in this paper we distinguish between attribute and property in the following way: we use the term attribute to designate a characteristic of a resource (e.g., author, first name,…), while the term property is used to designate a RDF property (e.g., http://purl.org/dc/terms/creator, http://xmlns.com/foaf/0.1/firstName,…). In simple terms, we call property an RDFized version of an attribute.

3 Research Design

The aim of this research was to propose a metadata application profile for technical reports in the context of Semantic Web and SDR. The research started with a literature review to identify appropriate metadata schemas and application profiles for technical reports/grey literature. From September to December 2018, bibliographical searches were carried out in order to identify MAPs for technical reports. We used the Scopus and Web of Science databases and the GoogleScholar and Google search engines using the following terms: “metadata schema”, “metadata vocabulary”, “application profile” and “dictionary metadata” associated with “technical report” and “grey literature”. The results of the search pointed to the absence of specific studies with that purpose. The search was also performed with the terms: “report or grey literature” and “representation”, “report or grey literature” and “access point”. We have identified studies about the metadata representation of grey literature, which includes the reports. However, these studies have not proposed a specific application profile [6, 9, 10, 15], which led us to create one.

The work of developing the AP for technical reports was carried out in four stages: (a) identification of the attributes needed to describe a technical report – mainly the study of Andrade, Shintaku and Barros [15]; (b) comparative analysis of data obtained by Andrade, Turner et al. [10] and Shintaku and Barros [15]; (c) data modeling; and (d) development of the MAP – this article reports on this part of the work.

We have based our work in the previous study of Andrade, Shintaku and Barros [15] as these authors have advanced in the identification of elements related to the descriptive and thematic representation of technical reports and by the gap of studies on the subject. Their study has identified 58 terms. We then analized and further reworked and arranged these terms, resulting in 7 classes and 60 properties, as showed in Fig. 1.

Although Dublin Core Terms is the most used metadata schema due to reasons widely presented in the literature, Andrade, Shintaku and Barros [15] found that its properties do not cover all technical report elements. The development of the data model showed the need to add properties that could reflect on the quality and coherence of the representation on the reports. Inclusions were carried out based on the authors’ experience, either in RDF and/or in Library Science.

The representation of the attributes identified by Andrade, Shintaku and Barros in a semantic web environment, in this case in scholarly digital repositories, requires the use of properties drawn from more than one metadata schemas. Conversely, schemas such as Marc21, which is widely used for technical report bibliographic description, have characteristics which makes their use difficult in our context, such as problems associated with cataloging rules (data structure) and the difficulty in representing hierarchies [16, 17]. And since there is no specific schema for technical reports that includes all the

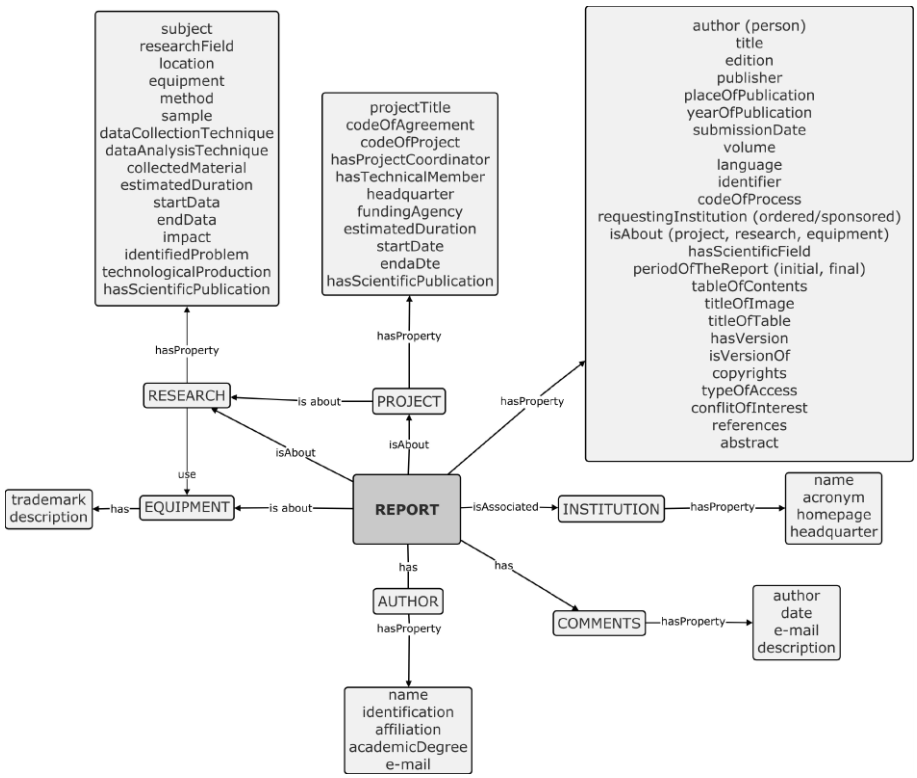

Fig. 1. Attributes related to technical reports (Source: Adapted from Andrade, Shintaku and Barros [15])

attributes identified by Andrade, Shintaku and Barros, the need to develop a metadata application profile is justified.

The MAP for technical reports, called TechRAP, aims to include metadata properties and rules that meet the needs of users of "a particular community with common application requirements" [18], namely, the community of SDR. For this purpose, Me4MAP was adopted [19].

Me4MAP is a method based on the Singapore Framework, which "[…] is the framework for designing metadata applications for maximum interoperability and for documenting such applications for maximum reusability" [20]. In this sense, the development of an application profile should include the following steps: identification of requirements; development of the domain model; and elaboration of the Description Set Profile (DSP) that, in this case, will be replaced by the elaboration of the constraints matrices. We have taken this option because DSPs are underused, even within DCMI, and the constraint matrices are more intelligible for humans.

After the definition of the detailed data model, Excel spreadsheets, called the Constraints Matrices, were elaborated, where information about cardinality and encoding schemes was included. Cardinality refers to specifying how many times a

metadata property may or may not appear in a description. The cardinality of a property may vary according to a particular context: a property may be mandatory in one profile and optional in another one. Or, a given property may be repetitive to allow more than one description or to allow different description options for a value. For example, when reporting an author's nationality, it is possible to repeat the property to say that (s)he has two or three nationalities, or repeat the property to encode the same value in two different ways: one using a string and another using an IRI.

4 TechRAP – Metadata Application Profile for Technical Report

The constraint matrix was elaborated based on the model proposed by Malta [19], but with two changes. In the header we include information about the respective class using the *rdf:type* property, and we include information regarding the Optionality, Min and Max, into a single column: Cardinality. The matrix includes the following information [19]:

(a) Label: property denomination;
(b) Property: the RDF property (the IRI);
(c) Range: the range of the property, where it is defined if the value is a Literal (L) or non-Literal (NL);
(d) Value string: is there a value string, yes or no?. The value string is "a Literal, optionally associated with either a syntax encoding scheme IRI or a value string language" [21];
(e) SES IRI: the IRI of the syntax encoding scheme, in case there is one. "The encoding scheme provides contextual information or parsing rules that aid in the interpretation of a term value". The syntax encoding scheme "indicates that the value is a string formatted in accordance with a formal notation" (for example, yyyy-mm-dd) [22];
(f) Value IRI: when the IRI identifies an actual value (yes/no). It is "an identifier that that identifies the actual value" [23];
(g) VES IRI: the IRI of the VES. The Vocabulary Encoding Scheme is a list of controlled term (Medical Subject Heading, Library of Congress Subject Heading). This is the identifier for term this list" [21, 23];
(h) Related description: to indicate (yes/no) if there is a related description, in this case a related matrix, and
(i) Usage: how the property may be used at instance level.

In the elaboration phase of the matrix, based on our expertise, some properties not mentioned in the study by Turner et al. [10] or Andrade, Shintaku and Barros [15] were added because they contribute to the consistency and quality of reporting. They are: *techrap:scheduledActivities, techrap:outcomesAnalysis, techrap:researchFieldActivity, techrap:guidelines, techrap:conclusion, techrap:hasIssue, schema:result, schema: scheduledTime, schema:reportNumber, onco:sampleSize, bibo:status*, *bibo:pages, frapo:hasGrantNumber, frapo:hasProjectIdentifier, dcite:geoLocation, dcterms:type, dcterms:format, dcterms:audience, and dcterms:source.*

Once this was done, we proceeded with the vocabulary alignment, that is, we analyzed metadata schemes to identify properties corresponding to the attributes. To that end, we used the Linked Open Vocabulary (LOV), "a high-quality catalog of reusable vocabularies for the description of data on the Web" [24]. To select a particular class/property, three aspects were considered: (a) use according to context; (b) property widely used by the metadata community; and (c) number of datasets that use the vocabulary. It is worth noting that a class/property was created only when: (a) no class/property was identified for a given attribute(s); or (b) to avoid using a large number of metadata schemas.

The creation of new properties (16) occurred with the development of a new vocabulary called "techrap". The namespace "techrap" was created using RDFS (http://opendata.dsi.uminho.pt/publications/techrap) and stored at http://hdl.handle.net/1822/61308. The Resource Description Framework Schema (RDFS) "is a semantic extension of RDF. It provides mechanisms for describing groups of related resources and the relationships between these resource" [25].

As a result of this work, several vocabularies were used (Table 1), where 7 classes and 89 properties come from. Table 2 presents the Technical Report constraints matrice. All constraints matrices can be acessed at link: http://hdl.handle.net/1822/61402.

Table 1. The namespace used

Vocabulary	Namespace
bibo	http://purl.org/ontology/bibo/
dbo	http://dbpedia.org/ontology
dcite	https://schema.datacite.org/meta/kernel-4.1
dcterms	http://purl.org/dc/terms
fabio	http://purl.org/spar/fabio
foaf	http://xmlns.com/foaf/0.1
frapo	http://purl.org/cerif/frapo/
onco	http://opendata.dsi.uminho.pt/health/onco-schema
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema
schema	http://schema.org
techrap	http://opendata.dsi.uminho.pt/techrap

The constraints matrices are shown in consonance with results of studies that sought to identify properties that represent technical reports, as is the case of Gonzalez de Gomez & Machado's studies; Zou et al. [9], cited by Turner, et al. [10] and Andrade, Shintaku and Barros [15] as it seeks to identify properties that optimize the discovery and reuse of such resources.

Table 2. Constraints matrices

Description: Technical report			rdf:type=http://purl.org/spar/fabio/TechnicalReport					Use: A report of a technical nature		
Label	Property	Range	Value string	SES IRI	Value IRI	VES IRI	Related description	Cardinality	Type	Usage
Abstract	dcterms: abstract	L	Y	————	N	————	N	1..*	String	A summary of the resource
Author	dcterms:creator	NL	N	————	Y	https://orcid.org/	Author	1..*	IRI	The responsible for making the resource
Conflit of Interest	onco:conflit OfInterest	L	Y	————	N	————	N	0-1	String	Conflit of interest
Coverage	dcterms: coverage	NL	Y	————	Y	http://purl.org/dc/terms/TGN	N	0…*	IRI	The spatial topic of the resource
Date	dcterms: created	L	Y	http://purl.org/dc/terms/W3CDTF	N	————	N	1	Date	Date of creation of the resource
Date of publication	dcterms: issued	L	Y	http://purl.org/dc/terms/W3CDTF	N	————	N	1	Date	Date of formal publication of the resource
Date submission	dcterms:date Submitted	L	Y	http://purl.org/dc/terms/W3CDTF	N	————	N	1	Date	Date of submission of resource
Edition	bibo:edition	L	Y	————	N	————	N	1	String	The edition of a document
Final considerations	techrap: conclusion	L	Y	————	N	———	N	1	String	Final proposition of reasoning

(*continued*)

Table 2. (*continued*)

Description: Technical report			rdf:type=http://purl.org/spar/fabio/TechnicalReport					Use: A report of a technical nature		
Format	dcterms:format	NL	Y	————	Y	http://www.iana.org/assignments/media-types/	N	1..*	IRI	The file format
Funding code	frapo:hasGrant Number	L	Y	————	N	————	Project	0-1	String	A data property specifying the grant number of a grant provided by a funding agency
Guidelines	techrap: guideline	L	Y	————	N	———	N	0-1*	String	Recommended action
Has version	dcterms:has Version	NL	N	————	Y	————	N	0-1*	IRI	A related resource that is a version, edition
Identifier	dcterms: identifier	NL	Y	————	Y	————	N	1..*	IRI	An unambiguous reference to the resource within a given context
Interested Group	dcterms: audience	L	Y	————	N	————	N	1*	String	A class of entity for whom the resource is intended or useful
Is Version Of	dcterms:is VersionOf	NL	N	————	Y	————	N	0-1	IRI	A related resource of which the described resource is a version, edition, or adaptation
Language	dcterms: language	L	Y	http://purl.org/dc/terms/ISO639-2	N	————	N	1*	String	A language of the resource

(*continued*)

Table 2. (*continued*)

Description: Technical report			rdf:type=http://purl.org/spar/fabio/TechnicalReport					Use: A report of a technical nature		
Legend	schema:caption	L	Y	————	N	————	N	0…*	String	Legend of an illustration or table
Licence	dcterms:rights	NL	N	————	Y	http://creativecomons.org/	N	1	IRI	Rights held in and over the resource
Pagination	bibo:pages	L	Y	————	N	————	N	1	String	Page range of a document
Period of the report	dbo:period	L	Y	http://purl.org/dc/terms/W3CDTF	N	————	N	1	Date	Period initial and final of report
Process number	frapo: hasReference Number	L	Y	————	N	————	N	0-1	String	A reference number for that item
Publisher	dcterms: publisher	NL	Y	————	Y	http://id.loc.gov/authorities/names.html	N	1..*	IRI	An entity responsible for making the resource available
Purpose	techrap:isAbout	NL	Y	————	Y	http://vocabularies.coar-repositories.org/documentation/resource_types/2.0.draft/	N	1	IRI	The purpose of this report is to project, research or equipment
References	dcterms: references	NL	Y	————	Y	————	N	1	IRI	A related resource that is referenced, cited by the described resource
Report number	schema: reportNumber	L	Y	————	N		N	1	String	The number or unique designator assigned to a report by the publishing Organization

(*continued*)

Table 2. (*continued*)

Description: Technical report			rdf:type=http://purl.org/spar/fabio/TechnicalReport					Use: A report of a technical nature		
Requesting institution	techrap: requesting Institution	NL	Y	————	Y	http://id.loc.gov/authorities/names.html	Institution	1*	IRI	Name of requesting institution
Research Field	techrap:research Field	L	Y	————	N	————	N	1	String	Research domain
Status	bibo:status	L	Y	————	N	————	N	1	String	Status of a resource
Subject	dcterms:subject	NL	Y	————	Y	http://id.loc.gov/authorities/subjects.html	N	1..*	IRI	The topic of the resource
Summary	dcterms: tableOfContents	L	Y	————	N	————	N	1	String	A list of subunits of the resource
References	dcterms: references	NL	Y	————	Y	————	N	1	IRI	A related resource that is referenced, cited by the described resource
Technical team	dcterms: contributor	NL	N	————	Y	https://orcid.org/	N	1..*	IRI	An entity responsible for making contributions to the resource
Title	dcterms:title	L	Y	————	N	————	N	1..*	String	A name given to the resource
Type	dcterms:type	NL	N	————	Y	http://vocabularies.coar-repositories.org/	N	1	IRI	The nature or genre of the resource
Type of access	dcterms:access Rights	NL	N	————	Y	http://vocabularies.coar-repositories.org	N	1	IRI	Declares the degree of openness of a resource
Volume	bibo:volume	L	Y	————		————	N	1	String	Volume of a resource

Note: NL – Non Literal, L – Literal, Y – Yes, N – No

It is believed that in this way, the contributions resulting from the field of activity of these researchers may contribute greatly to the description of technical reports. As these contributions are incorporated in the form of metadata properties and described by an application profile, the chances of optimizing the discovery and reuse of these resources is increased.

5 Conclusions

This article reports on the procedures for the development of a MAP for technical reports (TechRAP) in which classes and properties have been identified that may contribute to the discovery of this type of documents and optimize semantic interoperability for use and reuse. As a result seven constraints matrices were developed, holding 7 classes and 89 properties with respective encoding schemes, when they exist. Although there are metadata schemes such as Marc21, DC Terms or Funding Research Administration and Projects Ontology (Frapo) that can be used to represent technical reports, the use of the a single schema is not enough to describe specific aspects of these reports. By adopting different schemas to enrich the representation of the document and by using MAP, it is expected to ensure interoperability and semantics between information systems and so increase findability the resources in a domain-specific collection.

Currently, it is expected that this study may contribute to the visibility of this type of literature and also to related initiatives in the context of Semantic Web.

As future work we suggest the validation of the MAP and its test in the production environment and the development of ShEx schemas for data validation purposes.

We look forward to the outcome of the work of the DCMI Application Profiles Interest Group in order to apply it in TechRAP.

Acknowledgment. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019 and Universidade Federal do Espírito Santo.

References

1. Associação Brasileira de Normas Técnicas: Informação e documentação: relatório técnico e/ou científico. Apresentação. ABNT, Rio de Janeiro (2015)
2. Universidade Federal do Paraná: Relatórios. Ed. UFPR, Curitiba (2000)
3. Campello, B.L: Relatórios técnicos. In: Campello, B.S., Cendón, B.V., Kremer, J.M. (Org.) Fontes de informação para pesquisadores e profissionais. Ed. UFMG, Belo Horizonte (2000)
4. Kremer, C.D., Kovaleski, J.L., De Carvalho, H.G.: A memória técnica organizacional como prática de gestão do conhecimento em uma indústria metalúrgica de médio porte. In: Anais KMBrasil do Congresso Brasileiro de Gestão do Conhecimento (2008)
5. Abeckaer, A., Bernardi, A., Hinkelmann, K., Kühn, O., Sintek, M.: Toward a technology for organizational memories. IEEE Intell. Syst. **13**, 40–48 (1998)
6. González De Gómez, M.N., Machado, R.: A ciência invisível: o papel dos relatórios e as questões de acesso à informação científica. DataGramaZero - Rev Cienc. Inf. **8**(5) (2007)

7. Población, D.A.: Literatura cinzenta ou não convencional: um desafio a ser enfrentado. Ciência da Informação **21**, 243–246 (1992). http://revista.ibict.br/ciinf/article/view/438
8. Directory of Open Access Repositories: Content types in OpenDOAR repositories-Worldwide. University of Nottingham, Nottigham (2018)
9. Zou, X., Siyi, X., Li, Z., Jiang, P.: Constructing metadata schema of scientific and technical report based on FRBR. Comput. Inf. Sci. **11**(2), 34–39 (2018)
10. Turner, A.M., Liddy, E.D., Bradley, J., Wheatley, J.A.: Modeling public health interventions for improved access to the gray literature. J. Med. Libr. Assoc. **93**(4), 487–494 (2005)
11. Baker, T., et al.: Principles of metadata registries: a white paper of the DELOS Working Group on Registries (DELOS Network of Excellence on Digital Libraries) [s.d.], https://pdfs.semanticscholar.org/01ea/e200c915fbb38faf2584e87230bb15d2d683.pdf
12. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. Ariadne **25**, 27–31 (2000). http://www.ariadne.ac.uk/issue25/app-profiles/
13. Hillmann, D.: Using Dublin Core. Metadata Dublin Core® Metadata Initiative (2005). http://dublincore.org/documents/usageguide/. Accessed 03 Feb 2019
14. Baker, T.: A common grammar for diverse vocabularies: the abstract model for Dublin core. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005. LNCS, vol. 3815, p. 495. Springer, Heidelberg (2005). https://doi.org/10.1007/11599517_71
15. Andrade, M.C., Shintaku, M., Barros, P.P.: Proposta de elementos de metadados para representação e recuperação de memória técnica: o caso da Rede Ufes-Rio Doce. Cad. BAD **1**, 41–58 (2018)
16. Coyle, K.: Future considerations: the functional library systems record. Library Hi Tech. **2** (22), 166–174. https://doi.org/10.1108/07378830410524594
17. Fusco, E.: Aplicação dos FRBR na modelagem de catálogos bibliográficos digitais. Cultura Acadêmica, São Paulo (2011)
18. Lynch, C.A.: The changing role in a networked information environment. Libr. Hi Tech. **15** (1/2), 30–38 (1997). https://doi.org/10.1108/07378839710306981
19. Malta, M.C.: Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web Semântica (Tese). Universidade do Minho, Guimarães (2014)
20. Nilsson, M., Baker, T., Johnston, P.: The Singapore Framework for Dublin Core Application Profiles, 14 January 2008. http://dublincore.org/documents/singapore-framework/. Accessed 11 Feb 2019
21. Powell, A., Nilsson, M., Naeve, A., Johnston, P., Baker, T.: DCMI Abstract Model. DCMI. http://www.dublincore.org/specifications/dublin-core/abstract-model/
22. Dublin Core Metadata Inititative. DCMI grammatical principles (20017). www.dublincore.org/specifications/dublin-core/grammatical-principles
23. Coyle, K.: DCAM explained (2008). http://kcoyle.net/dcam.html
24. Vandenbussche, P.-Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semant. Web **1**, 1–5 (2014). http://www.semantic-web-journal.net/system/files/swj1127.pdf
25. W3C: RDF Schema 1.1. W3C recommendation, 25 February 2014. http://www.w3.org/TR2014/REC-rdf-schema-20140225