ISSN 1678-765X 10.20396/rdbci.v20i00.8666263

Article

RDBCI
Revista Digital de Biblioteconomia e Ciência da Informação
Digital Journal of Library and Information Science

Metadata standards in web archiving technological resources for ensuring the digital preservation of archived websites [1]

Danilo Formenton [1]; Luciana de Souza Gracioso [2]

ABSTRACT

Introduction: Digital preservation in Web archiving will only be possible with the effective use of metadata standards. These standards are the ones that determine the persistence, consistency, comprehensibility, the access, and representation of selected sites, collected and stored in Web archives, besides defining the archivability of sites and the interoperability among systems. Objective: In this context, the objective of the article was to identify and define which metadata standards could be judged by memory institutions and universities so that they could enable digital preservation in Web archives. Methodology: For this, a qualitative, exploratory, and descriptive research was done, using the bibliographic method from a non-systematic inventory together with a review and analysis of the literature content. The Dublin Core, MODS, EAD, VRA Core, PREMIS, and METS standards were selected and analyzed. Results and Conclusion: The analysis of the results indicates that Dublin Core, MODS, EAD, and VRA Core supported METS and PREMIS in detecting and documenting technical aspects of sites and proving their authenticity, context, and origin. METS can manage archived sites by acting as OAIS information packages, while Dublin Core proved to be an exponent for Web archiving through its use in remarkable area initiatives.

KEYWORDS

Digital preservation. Web archiving. Preservation metadata. Metadata standards. Information Science.

Padrões de metadados no arquivamento da web: recursos tecnológicos para a garantia da preservação digital de websites arquivados

RESUMO

Introdução: A preservação digital no arquivamento da *Web* só será possível com o uso efetivo de padrões de metadados, pois são eles que determinaram a persistência, a coerência, a compreensibilidade, o acesso e a representação de *sites* selecionados, coletados e armazenados em arquivos da *Web*, além de definirem a arquivabilidade de *sites* e a interoperabilidade entre sistemas. Objetivo: Neste contexto, foi objetivo do artigo identificar e definir quais padrões de metadados poderiam ser julgados por instituições de memória e por universidades

Author's correspondence

[1] Universidade Federal de São Carlos, São Carlos, SP, Brasil / e-mail: *formenton.danilo@gmail.com*

[2] Universidade Federal de São Carlos, São Carlos, SP, Brasil / e-mail: *lugracioso@yahoo.com.br*

[1] The article originates from a PhD Thesis under development, presenting changes in relation to the original text.

para que estas pudessem atender à preservação digital em arquivos da *Web*. Metodologia: Para isto, fez-se uma pesquisa qualitativa, exploratória e descritiva, que usa o método bibliográfico a partir de levantamento assistemático e de revisão e análise de conteúdo da literatura. Foram selecionados e analisados os padrões *Dublin Core*, MODS, EAD, VRA *Core*, PREMIS e METS. Resultados e Conclusão: A análise dos resultados aponta que *Dublin Core*, MODS, EAD e VRA *Core* amparam METS e PREMIS na descoberta e na documentação de aspectos técnicos dos *sites* e na comprovação de sua autenticidade, de seu contexto e de sua proveniência. O METS pode gerir *sites* arquivados, atuando como pacotes de informação OAIS, sendo que o *Dublin Core* mostrou ser um expoente para arquivamento da *Web* por seu uso em iniciativas notáveis da área.

PALAVRAS-CHAVE
Preservação digital. Arquivamento da Web. Metadados de preservação. Padrões de metadados. Ciência da Informação.

CRediT

• Recognitions: Not applicable.
• Financing: This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001.
• Conflicts of interest: Authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
• Ethical approval: Not applicable.
• Availability of data and material: Not applicable.
• Authors' contributions: Conceptualization, Data curation, Formal analysis, Funding acquisition, Research, Methodology, Project management, Resources, Supervision, Validation, Visualization, Writing - Original draft: FORMENTON, D.; Review & edit: FORMENTON, D; GRACIOSO, L. S.

JITA: JH. Digital preservation

Article submitted to the similarity system

Submitted: 03/07/2021 – **Accepted**: 16/08/2021 – **Published**: 11/01/2022

1 INTRODUCTION

Proposed by Tim Berners-Lee in 1989, the World Wide Web is a unique record of life in the 21st century and a unique information resource, hosting millions of websites, where different communities and individuals around the world connect (PENNOCK, c2013). However, not only does the web environment evolve at an intense speed, but information is also published and quickly moved to oblivion with the Internet and the abundant use of technologies. Likewise, live websites (and their webpages) are created quickly and their Uniform Resource Locator (URLs) and content regularly change and sometimes disappear completely, constituting complex, dynamic and ephemeral digital objects. All of this represents a very real threat to our individual, organizational, factual, or cultural digital memory, as well as to its technical legacy, evolution and social history (LIBRARY OF CONGRESS, [2021]; MASANÈS, c2006; PENNOCK , c2013; ROCKEMBACH; PEAVÃO, 2018). Due to this threat, according to Costa, Gomes and Silva (2017), organizations around the world – especially universities and cultural heritage institutions, such as libraries, archives and museums – have invested in policies, methods and technologies to collect, preserve over time and make accessible archived copies of web content.

In recent decades, digital preservation has become a subject of study that has established itself in Information Science. This is an emerging, collective and current problem in national and international productions in the area, requiring inter- and multidisciplinary analyzes and sustainable, integrated and collaborative solutions. One of the digital preservation strategies is the effective adoption of metadata standards to support the management, interpretation and preservation of digital objects in informational media, such as repositories.

Another notable strategy involves the digital preservation of website content. As a new topic in need of research and systematized initiatives in Brazil (ROCKEMBACH; PAVÃO, 2018), web archiving includes five steps, described in Kim and Lee (2007) and in Masanés (c2006), namely: selection (including preparation phases - defining the collection objective, capture policy and tools -, discovery - setting the entry points for capture, such as frequency and scope of capture - , and filtering for reduce the space opened by the previous phase to the limits in the selection policy), capture, archiving, access and quality review; this process may be extensive (non-selective), intensive (selective), topic-centered (thematic) and/or site domain. Linked to emerging Web archives, standards and guidelines for advancing and preserving the Web are created by consortia: the World Wide Web Consortium (W3C) and the International Internet Preservation Consortium (IIPC).

Integrated in web pages, in the form of links from one web page to others and user behavior records (RILEY, c2017), metadata (and metadata standards) have the function of describing only an informational resource in digital environments, multi-dimensioning its forms of access and use, ensuring their representation and recovery by the user. As an example, in the Web domain, the main standard is Dublin Core (DC); in the archival and museological domain, there is the Encoded Archival Description (EAD) and the Visual Resources Association (VRA) Core or the Categories for the Description of Works of Art (CDWA); and, in the bibliographic domain, we emphasize Machine Readable Cataloging (MARC) and Metadata Object Description Schema (MODS). For Formenton *et al*. (2017), metadata still defines the guarantee of preservation of a digital resource/object (for example, archived sites), through specific metadata standards, such as PREservation Metadata: Implementation Strategies (PREMIS) together with Metadata Encoding and Transmission Standard (METS).

Metadata standards, whether for description, management or preservation of digital objects, are key technological resources in interoperability. This function is ensured by practices and by description standards that translate into data encoding syntaxes, such as the Extensible Markup Language (XML) and the Standard Generalized Markup Language (SGML) Document Type Definition (DTD), in addition to the content standards (cataloging rules and codes), such

as Cataloging Cultural Objects (CCO), General International Standard Archival Description (ISAD(G)), Anglo-American Cataloging Rules (AACR2), and Resource Description and Access (RDA) and the value standards of data (vocabularies, thesaurus, and controlled lists), such as United States (US) Library of Congress Subject Headings (LCSH). These practices and standards are advocated by consortia, standards bodies and/or community leaders such as the Online Computer Library Center (OCLC), the International Organization for Standardization (ISO), the National Information Standards Organization (NISO) and the W3C.

Given the lack of national Web archiving studies that investigate, systematize and analyze in depth the metadata and the characteristics of the metadata standards applicable in the preservation of Web content in digital archiving systems, the need to identify and determine which metadata standards and schemas could be judged by organizations – especially cultural heritage institutions and universities – that are creating their systems, so that they could address digital preservation in Web archives. degree Metadata standards in the context of digital preservation and Web archiving have been discussed by Information Science and related areas, pointing out the elements of metadata that could be useful to the demands of structuring Web file systems in a more apt way the preservation of websites for historical, cultural and research purposes.

For this, a qualitative research is carried out, with an exploratory and descriptive approach (GIL, 2010; SILVA; MENEZES, 2005), which adopts the bibliographic method (MARCONI; LAKATOS, 2017; SEVERINO, 2016) from an unsystematic and of a review of the specific national and international literature of the last twenty years, directed and referring to standards and metadata schemas applied to digital preservation and archiving of the Web. Google Scholar and Scientific Electronic Library Online (SciELO) and Scopus and ScienceDirect (Elsevier), Emerald Insight (Emerald Publishing), Web of Science (Clarivate Analytics), Library & Information Science Abstracts (LISA) (ProQuest) and Library, Information Science & Technology Abstracts with Full Text (LIST) and Information Science & Technology Abstracts (ISTA) (EBSCO) available on the CAPES Journal Portal, plus site analysis, reports and guides of consortia and standardization bodies and/or community leaders, a definition, categorization and functions of metadata were recognized and systematized; the concept of preservation metadata and the information documented by metadata that supports long-term digital preservation management and Web archiving; and the main metadata standards used in the description and digital preservation of archived web content.

Thus, the present work is willing to expose the results and the analysis of the collected contents, and the product of this mapping predicted to collaborate with probable delimitations of guidelines and policies, which will be used by institutions interested and/or involved with the capture, the retention and permanent access to a website or archived site collection.

2 DEFINITION, CATEGORIZATION AND FUNCTIONS OF METADATA

As information created, saved and shared to describe objects, metadata allows us to interact with them to obtain the knowledge we need. Pervasive in information systems, metadata appear in various forms that show us how they are all structured to some extent, collected to serve a useful purpose and arranged in known categories. In the broad and classical definition that metadata means “data about data”, it is to be expected that metadata can be found anywhere, and it really is (RILEY, c2017). However, this literal and minimalist definition of the term metadata is not satisfactory, since, based on Alves (2017) and Sayão (2010), it is inexpressive and shallow given the complexity of the functions assigned to them in current information management context and also because it is necessary to understand them in the application domain where they are inserted. In this work, we will adopt the definition of metadata by Alves (2010, p. 47), as we consider it applicable to the Web domain and to specific

domains, such as the bibliographic domain, in addition to meeting the purposes of this investigation and being based on the construction standardized and consistent representations of univocal informational resources in different structured digital environments. In this way, metadata (metadata) can be conceptualized as:

> [...] descriptive elements or coded referential attributes that represent their own characteristics or those attributed to entities [...] in order to uniquely identify an entity (information resource) for later retrieval. (ALVES, 2010, p. 47).

For the author, the existence of metadata occurs through its encoding in standardized description structures called metadata standards (metadata statement), and the set of metadata or metadata elements (element sets) will integrate the metadata schema (metadata schema) of the metadata format or standard. According to Castro (2012) and Zeng and Qin (2008), the metadata element (metadata element) corresponds to a formally defined term to describe one of the properties (or attributes) of the resource of a certain type or with a particular purpose, such as 'the format' of a file.

In addition to the set of metadata (or prescribed elements, which are specified through statements), the metadata schema is composed of the value spaces, that is, the set of values and specification rules for each element and position in the descriptive structure, which are defined by standards external to it, as a syntax for expressing the values in elements, and coding schemas that fix coding rules, data syntax, and accepted forms/values. Such components will indicate the structural aspects (disposition of attributes and relationships between elements), syntax (coding of elements and logical order of values) and semantics (attribute meaning, etc.) for the definition of the pattern's metadata schema (ALVES, 2010; ZENG; QIN, 2008).

To better understand the concept of metadata, it is helpful to separate metadata into distinct categories that reflect key aspects of their functionality in a system. The main types of existing metadata are used under the particularities of the domain (and the functions to be performed), the demands of users and the types of resources/entities for representation (ALVES, 2017; GILLILAND, c2016; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004). From the Digital Preservation Coalition ([201-?]), Riley (c2017) and Sayão (2010) are considered several functional categories of metadata types, thus being understood:

- Descriptive Metadata – detail a digital resource for location, identification or understanding. They may include properties or elements such as title, author, and subject, where primary uses are discovery, presentation, and interoperability.
- Structural Metadata – they explain the internal structure of the digital archive and the hierarchical relationships of the resources that are part of each other. They can have properties such as order and place in the hierarchy, where the primary uses are navigation and presentation.
- Administrative Metadata – they provide information that supports the lifecycle management (creation, selection, description, etc.) of information resources. These may include properties such as file type and size, creation date/time, preservation event, copyright status, and license terms, where primary uses are interoperability, digital object management, and preservation. These are subdivided into:
  - Technical Metadata – indicate the technical aspects and dependencies of a digital file to decode and render it.
  - Preservation Metadata – include information (e.g. hardware and software dependencies) required for long-term management of a digital archive.
  - Copyright Metadata – document information to support the management of intellectual property rights associated with a content.
- Markup Languages – include metadata and flags for other structural or semantic

features in the content. May contain properties such as paragraph, name, list and date, where primary uses are navigation and interoperability.

In view of this, a notable reason for creating descriptive metadata is to make it easier to discover relevant informational resources in the Web domain or in specific domains; in addition, metadata can help organize electronic resources, promote interoperability, support archiving and preservation, in addition to other common activities to be done in a digital information system, which, as stated by Gilliland (c2016) and the National Information Standards Organization (c2004), depict some of the primary functions of metadata. For the purposes of this work, we will highlight the preservation metadata class as they are vital for achieving effective management and for the long-term preservation of digital and electronic files; and the class of descriptive metadata, the best-known facet of metadata (SAYÃO, 2010), which, based on the indications of the international working group Web Archiving Metadata (WAM) of OCLC Research by the studies by Dooley and Bowers (c2018), Samouelian and Dooley (c2018) and Venlet *et al*. (c2018), will be addressed in accordance with best practices for creating coherent and efficient descriptive metadata about archived web content (or rather websites) and for Web archiving.

3 PRESERVATION METADATA AND DESCRIPTIVE METADATA FOR WEB ARCHIVING

Considering that digital preservation is a management process, preservation metadata is primarily categorized as administrative metadata, but it is permissible that preservation metadata schemas include elements that span multiple categories such as descriptive, structural, and administrative. Such metadata make up a crucial part of digital preservation strategies and are conceived in the PREMIS data dictionary (a de facto international standard for preservation metadata) (CHEN; REILLY, 2011; DAPPERT *et al*., 2013; SAYÃO, 2010). *Premis Editorial Committee* (2015, p. 2, our translation) defines preservation metadata "[...] as the information that a repository uses to support the digital preservation process". According to Dappert and Enders (2010) and Caplan (2017), this is information that describes a digital resource in the repository to ensure its access and long-term use. For Márdero Arellano (2008), they refer to the content of the resource, its context and the structure of creation, in addition to the changes made in its life cycle. Defined like this, it is noted that preservation metadata are built to fulfill a wide range of different but related functions (SAYÃO, 2010). These metadata support the different requirements of digital preservation which, according to Lavoie and Gartner (c2013) and Premis Editorial Committee (2015), intend to maintain availability; rendering (making the object noticeable to a user via reproduction – for visual materials –, display – for audio materials –, or by other means inherent to its format); understandability; the identity; persistence; authenticity (the quality that the object is what it intends to be, with which the integrity of its content and origin can be verified); and the viability (property of being readable by the storage media) of digital objects for long periods of time.

That said, *Digital Preservation Coalition* ([201-?]) and Gilliland (c2016) deduce certain reasons why metadata is important for digital preservation, which, together with the considerations of the aforementioned authors, can be described as follows:

- Decision making – information linked to a digital object, such as the software to open it, the time it needs to be kept, or the history of changes made to it, help professionals make decisions about how and why to preserve it..
- Legal aspects – metadata allow systems to track existing rights levels, licenses, and reproduction information for the original items, their associated objects, and multiple

versions of these.

- Subsistence – metadata documentation of how the information object was created and maintained, how it behaves and how it links to other objects will be crucial to its existence, regardless of the current system used to store and retrieve it.
- Meaning context – metadata provide context information required for future users to understand the meaning of the content of a record, playing a vital role in documenting relationships and in indicating authenticity, structural/procedural integrity and the degree of completeness of objects.

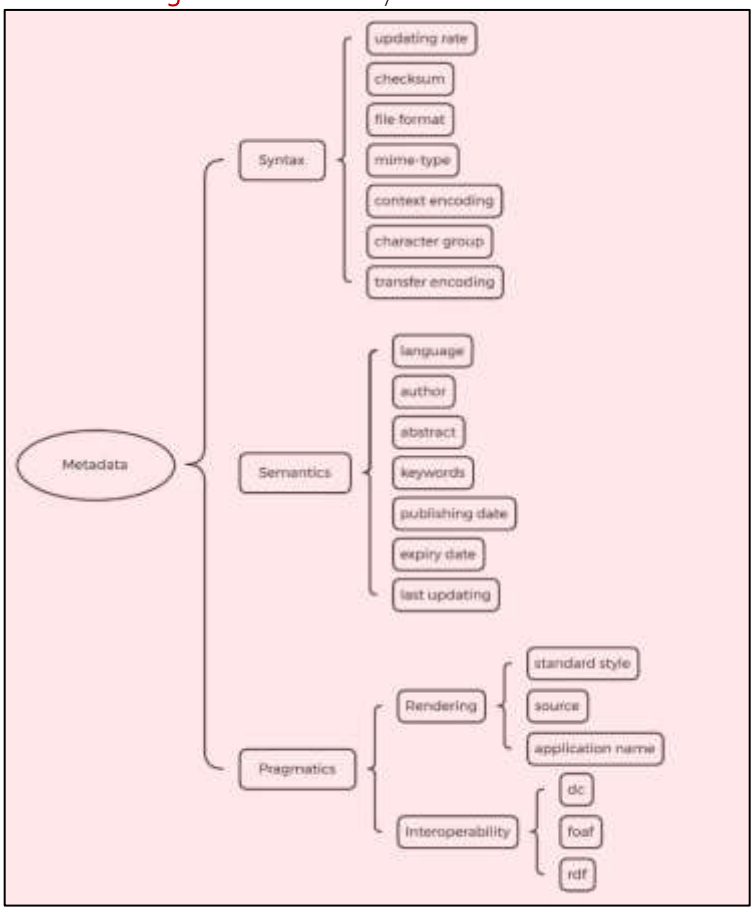
These justifications express certain descriptive, administrative and structural information to be incorporated by preservation metadata. In this sense, grouping the weightings of Caplan (2017), of Dappert *et al*. (2013), by Dappert and Enders (2010), by Formenton *et al*. (2017), by Lavoie and Gartner (c2013), by the National Library of New Zealand (2003) and by Sayão (2010), we identified a set of information and interrelated functions that support the management of digital preservation, covered in the capture, creating and maintaining preservation metadata:

- Recording information about agents - people, organizations, software and hardware - with functions in rights, in computational environments and in actions - preservation, dissemination, access, use, etc. – that affect the object.
- Recording of technical dependencies needed to access, render – or present, execute etc. – and use the object.
- Recording information that establishes the significant properties of the object, that is, characteristics of the original object and the environment that must be maintained by preservation actions for a community of users (for example, images on a web page), guiding decisions about which actions should be selected.
- Recording of the object's physical and logical structural relationships (for example, which image is integrated into which website and which page follows which in a digitized book), as well as information about its storage medium.

Although there is little work that brings together and synthesizes experiences of implementation of preservation metadata for the accumulation and consolidation of best practices in digital preservation, or even that assesses the costs involved in collecting and managing preservation metadata and the practical benefits In addition to incurring these costs (LAVOIE; GARTNER, c2013), preservation metadata is a key component of all digital archiving. Such metadata document information on content and provenance, authenticity, fixity, reference, context, rights, etc. aligned with the information model of the Open Archival Information System (OAIS) and its three information packages (i.e., Information Submission Package - PSI, Information Archival Package - PAI and Information Dissemination Package - PDI), the which ensure that digital resources/objects are maintained, retained, identified, accessed, deciphered, rendered and used cohesively and accurately over time.

As indicated by Banos *et al*. (c2013) and by Melo and Rockembach (2020), the use of metadata is also one of the main archivability facets of websites, or rather, factors that must be taken into account to calculate the extent to which the website satisfies the conditions for securely transferring your content to a web archive for preservation purposes. Based on a general model of shared perspective across different information disciplines – Philosophy, Linguistics, Computer Science, etc. – the authors consider metadata at three levels (summarized and shown in Figure 1) to measure a site's archival capacity (archivalability), namely: syntax (e.g., how this is expressed); semantics (for example, what it is about); and pragmatics (for example, what can be done with it).

Figure 1. Archivability Facets: Metadata.



Source: Adapted from Banos *et al*. (c2013).

Banos *et al*. (c2013) explain that content encoding and transfer metadata can be embedded by the server in Hypertext Transfer Protocol (HTTP) headers; in addition, rendering metadata such as the name of the app, the end-user language to understand the content; and descriptive information such as author and keywords that help you understand how content is classified can be included in the attribute and values of the HyperText Markup Language (HTML) element. To promote better interoperability, the authors recommend the use of known metadata and description schemas, such as DC, Friend of a Friend (FOAF) and Resource Description Framework (RDF); moreover, the existence of selected metadata elements is checked to increase the possibility of implementing automated extraction and refinement of metadata in the collection and ingestion of Web content or, soon after, in the repository management phase.

Judging that web archiving is a relatively new process for cultural heritage institutions, there are few standards; thus, metadata practices vary greatly among the different initiatives in the area, whether among national libraries or even due to differences in description approaches between the two traditions of bibliographic and archival description of resources that do not promote the interoperability of metadata, as , for example: in cataloging in libraries the nature of the content is revealed mainly by the title (if it is descriptive) and the subject terms; otherwise, archivists routinely use extensive free-text notes to describe both the content and context of the material (DI PRETORO; GEERAERT, 2019; DOOLEY; BOWERS, c2018). Of these metadata practices in web archiving initiatives, we mentioned:

- Arquivo.pt, a service of the Foundation for Science and Technology (FCT) of the Ministry of Education and Science of Portugal, which allows searching and accessing

archived Portuguese web pages, elucidates certain metadata about the contents of a website for its preservation, such as: Description (description), short text describing the content of the page; Keywords (keywords), expressions representing the main themes of the page; and Dublin Core, DC metadata (ARQUIVO.PT, 2018).

- The Internet Archive, a foundation that provides free and universal access to a digital library with more than 500 billion Web pages and other archived content, indicates metadata that has special meaning in describing the content of the items in the archive, such as Sponsor (sponsor); Scanner; Scan Date (scandate); Image Count (imagecount); and Media Type (mediatype) (INTERNET ARCHIVE, 2018). In fact, as cited by Samouelian and Dooley (c2018), their web archiving service has sixteen DC fields that users can choose from, and the ability to manually add custom fields.
- In the Library of Congress Web Archive, a program that manages, preserves, and grants access to archived Web content, sub-elements of MODS metadata are encoded in the register of sites for thematic and event-based collections, such as: Text (<text>) for scopes (domain), in the element Part (<part>); Identifier (<identifier>) for the URL of the source, in the Related Item (<relatedItem>) element; and Place (<place>) in the Source Information element (<originInfo>) (LIBRARY OF CONGRESS, [2021]).

Therefore, according to Dooley *et al*. (2017), OCLC Research established the WAM working group in the face of the challenge of the lack of a common approach to creating metadata in the web archiving community. , providing a bridge between bibliographic and archival approaches to description, the group has published three reports that include: a literature review of the descriptive metadata needs of end users of Web archives and the professionals who create and manage such metadata (VENLET *et al*., c2018); an analysis of Web collection tools, with a view to their functionality for extracting descriptive metadata from the tracked files (SAMOUELIAN; DOOLEY, c2018); and guidelines to help institutions and individuals improve the consistency and efficiency of their metadata creation practices in this emerging area (DOOLEY; BOWERS, c2018).

In this latest report, the WAM group indicates a lean set of data elements, with content definitions and usage notes (i.e., a data dictionary) suited to the unique characteristics of archived websites relevant to the description of materials in libraries and files, such as at the item and collection levels, which can be used alone or together with other more granular content and data structure standards (DOOLEY *et al*., 2017; DOOLEY; BOWERS, c2018). Furthermore, according to Di Pretoro and Geeraert (2019), each WAN data element contains the advantage of brief mappings (crosswalks) for the DC, EAD, MARC 21, MODS and schema.org, which are intended to facilitate such conversions. Based on Dooley and Bowers (c2018), the set of fourteen WAM data dictionary data elements for describing websites or archived site collections is composed of:

1. Collector – the institution in charge of curating and managing an archived site or collection.
2. Contributor – the entity (organization or person) that has made significant but minor contributions to the content of a site or archived collection.
3. Creator – an organization or person with primary responsibility for having created the intellectual content of an archived site or collection.
4. Date – a single date or date range linked to an event in the lifecycle of a site or archived collection.
5. Description – notes that explain the content, context, and aspects of a site or archived collection.
6. Extent – an indication of the size of a website or archived collection.
7. Genre/Form – a term that determines the type of content for a site or archived collection.

8. Language – the language(s) of the archived content, including visual and audio resources with linguistic components.
9. Relation – the whole/part relationships between a single archived website and any collection to which it belongs.
10. Rights – declarations of rights and legal permissions granted by intellectual property rights or other legal agreements.
11. Source of description – information about extracting/creating the metadata itself, such as data sources and date of obtaining data from the sources.
12. Subject – the main topic(s) describing the content of a site or archived collection.
13. Title – the name by which a site or archived collection is known.
14. URL – the Internet address of a website or archived collection.

Based on another approach, by analyzing the metadata of various Web archiving projects, such as the National Library of Australia and the Smithsonian Institution Archives in the United States, Kim and Lee (2007) suggest descriptive and administrative metadata for intensive Web archiving. Judging that most of the revised project metadata is DC-based and that intensive web archiving requires more detailed metadata elements due to quality-oriented selectivity, the authors adopted other common administrative elements in addition to the basic DC elements in these projects, like:

- Availability – how web content can be obtained or contact details.
- Audience – the expected group to use web content.
- Date captured – the date associated with the capture of the site in the file.
- Date validated – the date the webpage was validated, as being actually encoded, using the W3C Markup Validation Service or other services.
- Collecting method – the method of collection of web content, such as automatic, manual, or downloaded.
- Collecting tool) – the software needed in the Web content collection process.

Indeed, the DC schema proves to be remarkable for the description of archived Web content, given the substantial similarities of the standard with the WAN data element set (DOOLEY; BOWERS, c2018), the adoption and adaptation of its essential elements in the metadata from web-intensive archiving initiatives, or the simplicity of the pattern that motivates its general use in web-intensive archiving (KIM; LEE, 2007). Despite the ambiguity involved in the scope of preservation metadata, which, according to Dappert and Enders (2010) and Lavoie and Gartner (c2013), is portrayed by the difficulties in accurately categorizing them, which can extend across all metadata classes, the work focused on descriptive and preservation metadata that support the discovery, identification, presentation, interoperability, and long-term digital preservation of archived site collections..

In this sense, the definition and, perhaps, the adaptation of metadata standards and schemas make it necessary to act on digital preservation policies in Web archiving. The stages of the archiving process (selection, capture, etc.) should be considered, technologies (tracking robots, etc.) and the archiving methods adopted (domain, theme, etc.), and the types of web content collected, maintained and made available (web page, social network, etc.), as well as meeting the needs of end users, the series of information to be registered and the decisions made regarding copyright, privacy, cost, quality, etc. Issues, and a future of unpredictability inherent in the digital preservation of information published on the Web.

4 IDENTIFICATION OF PATTERNS AND METADATA SCHEMAS FOR WEB ARCHIVING

The usefulness of metadata comes from its understandability by software applications and by the people who use them. Known as metadata vocabularies, sets of elements or also as formats, schemas can be formally standardized through standardization organizations (ISO, NISO and W3C, for example) and, in addition, hosted and maintained by bodies industry or community leaders, such as the US Library of Congress, who endorse them for use in their target communities (RILEY, c2017). Metadata standards help make metadata as useful as possible because, according to the Digital Preservation Coalition [201-?], they provide guidelines for uniform formatting as schemas are guidelines for uniform metadata formats, so, standards and schemas ensure that metadata for digital records is interoperable.

That said, also called schemas, metadata schemas are the set of metadata elements (and rules for their use) of a standard created for a purpose, such as describing a type of informational resource (CHAN; ZENG, c2006; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004). In Zeng and Qin (2008, p. 323, our translation), metadata schemas constitute:

> A machine-processable specification that defines the structure, syntax encoding, rules, and formats for the set of metadata elements in a formal language in a schema. In the literature, the term metadata schema usually refers to the set of elements in their entirety, as well as the coding of elements and structure with a markup language.

In fact, through Castro (2012), Chan and Zeng (c2006), National Information Standards Organization (c2004) and Vellucci (2000), it appears that the schema (schema) is an entity as a whole, including the semantic components and of content (called a set of metadata elements), such as encoding the metadata with a syntax or markup language (the MARC format and an XML/SGML DTD, for example), which have three basic parts or characteristics:

1. Structure – the data model or architecture used to hold the metadata and the way the metadata statements are expressed. As examples, we can mention the RDF metadata architecture and the XML METS schema.
2. Semantics – the names and meanings of the elements and their refinements.
3. Contents – the statements or instructions of how and what values should be assigned to the elements.

Thus, the metadata schema defines attributes and rules, under semantic and structural aspects, consisting of other types of schemas (schemas), which determine the syntax of data encoding, which in turn helps to establish the structure and semantics (meaning) of attributes and values in a metadata standard (ALVES, 2010; ZENG; QIN, 2008). From the literature review and analysis, we identified several patterns and metadata schemas used to describe resources in different domains. Most recurring patterns have their origins when the Web was in its infancy. In the second half of the 1990s and early 2000s, there was a rapid development of formats for the needs of specific communities and the encoding of complex digital objects, which are delimited by their own sets of metadata elements, their particularities, and the application domains. Some of the main metadata standards in force and indicated in the specialized literature for Web archiving are discussed below when the mapping and indication of elements for Web archiving are exposed.

4.1 Dublin Core Standard

The DC began in Chicago, at the 2nd International World Wide Web Conference, in 1994, in a debate on semantics and the Web in view of the difficulty of discovering resources. This fact led OCLC and the National Center for Supercomputing Applications (NCSA) to hold the OCLC/NCSA Metadata Workshop in the North American city of Dublin, Ohio, in 1995, where they discussed how a basic semantic set would be useful for the search and the retrieval of web-based resources. The result was called "Dublin Core metadata" based on the location of the workshop (DUBLIN CORE METADATA INITIATIVE, c2020a). According to Harper (2010) and Sayão (2010), the set of DC elements is small and simple, so that it is semantically understandable; moreover, the DC is represented by various syntaxes, such as coded in HTML or XML and structured in RDF, providing exchange and reuse.

Today, at version 1.1, DC is a two-level vocabulary: simple and skilled. Thus, the simple DC comprises fifteen essential properties or elements (the core) and the qualified DC contains additional elements, in addition to qualifiers that specify the meaning of the element (element refinement) or identify schemes in interpreting its value (coding scheme) (DUBLIN CORE METADATA INITIATIVE, 2012, 2020b). For the scope of digital preservation, Formenton *et al*. (2017) highlights some qualified DC elements, such as, for example, Format (format), Identifier (identifier), Rights (rightsHolder) and Provenance (provenance), which, although more focused on access than for preservation, record information provided for in PREMIS preservation metadata.

Regardless of the criticisms of the structure and the very simplistic and generic set of DC elements (above all, compared to other formats, such as MARC), DC is a guide for semantic interoperability and consensus among different communities in the world, and it even plays a role of leadership in creating descriptive Web archiving metadata (DOOLEY *et al*., 2017; DOORN; TJALSMA, 2007; HARPER, 2010). As pointed out by the WAM working group at OCLC Research, as per Dooley and Bowers (c2018), Samouelian and Dooley (c2018) and Venlet *et al*. (c2018), DC Descriptive Metadata Schema in Version 1.1 is widely used in describing archived websites by users of Archive-It, an Internet Archive subscription web archiving service.

4.2 MODS Standard

Designed by the US Library of Congress in 2002, the MODS scheme can be adopted in particular for library applications. Expressed in XML, this descriptive metadata standard includes a subset of MARC 21 fields and uses word-based rather than numeric labels, allowing for easy understanding (LIBRARY OF CONGRESS, 2016, 2018). As advantages of the MODS, according to Guenther (2003) and Mccallum (2004), it is noted that the MODS is simpler than the complete MARC and provides a richer description compared to qualified DC; moreover, in MODS there is the regrouping of certain MARC elements and, in some cases, what is in several MARC elements is merged into a single MODS element. For example, MARC fields and subfields for the main and minor entry of Name (100 and 700) are regrouped in the Name (<*name*>) MODS element.

Currently in version 3.7, the MODS schema has a set of twenty top-level descriptive metadata elements, through which it provides bibliographic information that integrates other XML schemas, such as METS and PREMIS. Under the focus of digital preservation, Formenton *et al*. (2017) look at three MODS elements: Source Information (<titleInfo>), Related Item (<relatedItem>), and Access Condition (<accessCondition>). For the authors, these elements document useful information that aids preservation metadata, whether in proving the authenticity, integrity and provenance of objects or in identifying the rights of the electronic resource that intervene in the preservation, access and use of its contents.

Although MODS elements inherit the semantics of MARC elements, converting an original MARC record to MODS and then back to MARC results in loss of data or some loss of specificity in the markup. In certain cases, if reconverted to MARC 21, the data may not be inserted in exactly the same field in which they started, as a MARC field may have been mapped to a more general one in MODS and, in view of this, the data itself will not be lost, only the detailed identification of the type of element they represent. In other cases, a MARC element (for example, field 340 Physical Medium) may not have an equivalent MODS element, and then specific data may be lost when converting to MODS. Therefore, MARC XML must be used before for a lossless exchange (GUENTHER, 2003; LIBRARY OF CONGRESS, 2016). On examples of the use of MODS in Web archiving, Guenther and Myrick (2007) point to the Library of Congress Web Archive, originally created in the project "Mapping the Internet Electronic Resources Virtual Archive" (MINERVA) in partnership with the Internet Archive, which is made up of archived site collections that are cataloged with MODS.

4.3 EAD Standard

The EAD scheme originated in a 1993 University of California, Berkeley library project. Directed by Daniel Pitti, the Berkeley project aimed to develop a non-proprietary coding standard for computer-readable research instruments such as inventories, indexes, records, guides and documents created by archives, libraries, museums and repositories to support the use of their collections (LIBRARY OF CONGRESS, 2013). According to Allison-Bunnell (2016) and Pala (2017), the EAD3 version focuses on simplifying the standard and increasing clarity and semantic consistency compared to the EAD 1.0 and EAD 2002 versions, promoting interoperability and improved functionality in international and multilingual environments.

Today, in EAD3 version 1.1.1, this XML standard has a set of one hundred and sixty-five descriptive elements and eighty-five attributes, which provides bibliographic information that aligns with other XML schemas, such as the Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC-CPF) (SOCIETY OF AMERICAN ARCHIVISTS, 2019). For the purpose of digital preservation, Formenton *et al*. (2017) note certain EAD 2002 elements maintained in version 1.1.1 of EAD3, such as the Archival Description (<archdesc>). According to the authors, the DC, MODS and EAD standards, even if they are more applicable to the discovery, search, retrieval or location of resources than to preservation, are useful schemes for recording descriptive metadata in support of PREMIS and to METS.

Although the lack of resources and knowledge/expertise available in an institution has influenced its adoption, in the last twenty years, as raised by Eidson and Zamon (2019), EAD remains relevant due to the large number of files that adopted it and continue currently using it to publish their research instruments online. Examples of using EAD in web archiving is the California Online Archive, which provides free, public access to detailed descriptions of primary source collections maintained by institutions across the state of California, such as the University of California Web Archive. California in Irvine https://oac.cdlib.org/findaid/ark:/13030/c8q81jn9/, through EAD research instruments.

4.4 VRA Core Standard

Developed by the VRA in 1996, the VRA Core is a blueprint for describing visual cultural works – including paintings, drawings, sculptures, architecture, photographs, etc. –, and images that document them. It is used as a stand-alone format and as an extension scheme

of METS[2] for objects that contain cultural heritage resources (VISUAL RESOURCES ASSOCIATION, 2014). According to Lima, Santos, and Santarém Segundo (2016) and Lubas, Jackson and Schneider (2013), this standard has versions, with VRA 1.0 (1996) based on CDWA, VRA 2.0 (1996), which indicated the search for CCO standards and VRA 3.0 (2000), which resembles DC in simplicity, number of elements and qualifiers.

Currently, in version 4.0, released in 2007, the VRA Core XML schema supports interoperability and exchange of records. VRA Core 4.0 has nineteen descriptive elements and nine global attributes, in which the top-level wrapper element – Work, Collection (collection) or Image (image) – includes the other eighteen elements in individual records (VISUAL RESOURCES ASSOCIATION, 2007, 2014). For digital preservation purposes, we note elements such as Location, Rights and Source, which may support PREMIS and METS in identifying and defining trustworthiness, authenticity, integrity, provenance and the context of cultural works and their representations.

Despite its specificity and the imposition of certain restrictions on the creation of links to non-VRA Core records, plus the fact that it is less common compared to other formats, as stated by Eito-Brun (2015) and Senander III (2013), it is It is possible to create links in the schema for search instruments and the process of converting VRA Core records to MARC records is quite simple, straightforward and efficient. As for the examples of using VRA Core 4.0 in web archiving, indirectly there is the Cornell University Library web archive which has collection sites cataloged with the VRA Core, such as Mysteries at Eleusi Images of Inscriptions and Billie Jean Isbell Andean Collection: Images from the Andes.

4.5 PREMIS Standard

PREMIS is named after a working group sponsored by OCLC and the Research Libraries Group (RLG) in the United States from 2003 to 2005. This group created a final report in 2005 called the Data Dictionary for Preservation Metadata, which defines a core set of semantic units distributed in four types of entities related to each other in their data model (Objects, Events, Agents and Rights), being implementable and of wide application, in order to support digital preservation in repository systems. In the PREMIS data dictionary, 'semantic unit' corresponds to a piece of information or knowledge and is the properties that describe important entities with roles regarding digital preservation activities, i.e., digital objects and their contexts, life cycle events , agents involved in preservation and rights. That said, 'entity' would be an abstraction for a set of "things" (environments, events, etc.) described by the same properties (CAPLAN, 2017; DAPPERT; ENDERS, 2010; PREMIS EDITORIAL COMMITTEE, 2015).

The PREMIS data dictionary does not target certain metadata classes that are already well met/supplied by existing standards, such as descriptive metadata and format-specific technical metadata, thus often combining with other different standards (METS, Metadata Authority Description Schema – MADS, Z39.87/NISO Metadata for Images in XML Schema – MIX, for example) to cover complementary functionalities supported by them. Furthermore, although heavily influenced by the OAIS model, which is widely accepted as one of the main standards to be followed to normalize digital preservation repositories, PREMIS provides key information that covers the entire lifecycle of digital objects and goes beyond the scope of the repository, as it provides specific information to preserve digital objects, while OAIS provides broader categories of this information; and allows the recording of information about digital objects that occur before being entered into the system, which is not covered by OAIS (GUENTHER; DAPPERT; PEYRARD, c2016; LAVOIE; GARTNER, c2013; SAYÃO, 2010).

2 External schemas for use with METS. Available at: https://www.loc.gov/standards/mets/mets-extenders.html. Access on: 23 July 2021.

Today, in version 3.0, issued in 2015, the PREMIS data dictionary presents guidelines for organizing and designing preservation metadata. As mentioned above, the PREMIS data dictionary is structured around a data model and also implementations, such as the associated standard XML schema[3], which define what "things" need to be described (the PREMIS entities) and what information they need to be known by the preservation repository to be told about them (the semantic units of PREMIS) (GUENTHER; DAPPERT; PEYRARD, c2016; PREMIS EDITORIAL COMMITTEE, 2015). According to Caplan (2017), this XML schema provided by PREMIS Maintenance Activity matches directly with the PREMIS data dictionary, allowing the description of Objects, Events, Agents and Rights, such as the use of PREMIS represented in XML for the exchange of metadata between systems preservation. For Formenton *et al*. (2017), because PREMIS applies the OAIS information model and the requirements for the preservation of digital objects (authenticity, provenance, etc.), all its entities/semantic units are vital to digital preservation.

Although the lack of training/expertise and integration with the existing system can bring barriers to its adoption in cultural heritage institutions (ALEMNEH; HASTINGS, 2010), the PREMIS data dictionary provides a remarkable framework for describing and preserving computational environments (hardware, software, etc.) that support the rendering or execution of digital objects and their long-term use. As an example, PREMIS is adopted in the description of rendering environments for web content from the National Library of France, which hosts the French web archive (DAPPERT *et al*., 2013). Other examples of using PREMIS in web archiving include Bailey and LaCalle (2015) and Rowell and Krewer (2016), who present the Internet Archive view of how PREMIS preservation metadata interacts with the ARCHive Web format (WARC), a file standard for web content.

4.6 METS Standard

Created by the Digital Library Federation (DLF), METS was pioneered by the 1997 Making of America II (MOA2) project, which developed an XML document format for encoding descriptive, administrative, and structural metadata for textual and image-based works. Expressed in XML, METS makes it possible to encode the metadata necessary for managing digital library objects in a repository and for exchanging these objects between repositories (or between repositories and their users) (LIBRARY OF CONGRESS, 2017). According to Cantara (2005), Mcdonough (2006) and Sayão (2010), METS is a flexible mechanism to organize all metadata associated with the digital object, express the complex links between multiple metadata classes and, additionally, associate an object with behaviors or services, supporting interoperability, scalability and long-term digital preservation.

Currently in version 1.12.1, 2019, the METS scheme is aimed at encoding complex objects in digital libraries. To share XML documents conforming to METS and establish common practices, the standard defines the components of a METS profile and the XML schema for encoding it. These profiles describe in detail a class of METS documents to create and process METS documents according to a specific profile, with a METS document comprising seven main sections: METS header (<metsHdr>), Descriptive metadata (<dmdSec>), Metadata Administrative (<amdSec>), File (<fileSec>), Structural Map (<structMap>), Structural Links (<structLink>) and Behavior (<behaviorSec>) (DIGITAL LIBRARY FEDERATION, 2010). In digital preservation, Formenton *et al*. (2017) point out that a METS document can act in the execution of OAIS information packages.

Although it is possible to use METS with PREMIS, this is not entirely simple for some reasons. There is an imperfect correlation between the two structures, as the first divides information into distinct sections, depending on whether it is technical, rights, etc. metadata, and the second has sections for Objects, Events, etc. Also METS and PREMIS have some

3 Available at: https://www.loc.gov/standards/premis/v3/premis-v3-0.xsd. Access on: 26 July 2021.

duplication (for example, each defines a tag for storing checksums, imposing the decision to register these duplicate elements in METS sections, PREMIS sections, or both), which implies the need for adoption best practices for using them together[4] in support of non-variation of data representation and promoting interoperability (CAPLAN, 2017). Furthermore, as mentioned by Lavoie and Gartner (c2013), the flexibility included in METS can cause interoperability problems, because when such diverse content, treated in various ways, is allowed within the sections of a METS document, it becomes more difficult to exchange METS records. However, this is mitigated to some extent by registered METS profiles[5]. For example, the METS profile for website captures from the ECHO DEPository project at the University of Illinois at Urbana-Champaign (HABING, 2006) aims at the transfer and digital preservation of web capture content between repositories. Other examples within the web archive include Truman (2016) and Veikkolainen and Lager (2016), who exhibit the Finnish web archive maintained by the National Library of Finland, where the content comprises files in WARC format in METS information packages.

5 ANALYSIS OF METADATA PATTERNS IN LIGHT OF DIGITAL PRESERVATION IN WEB ARCHIVING

The DC, MODS, EAD, VRA Core, PREMIS and METS metadata standards and schemas have common characteristics and some peculiarities. The considerations carried out here were based on the principles of long-term digital preservation, the definitions of the OAIS information model, the information expressed by the preservation metadata and the descriptive and administrative metadata for Web archiving, especially with the WAN elements of Dooley and Bowers (c2018), elements from Kim and Lee (2007) and metadata from Banos *et al.* (c2013) and initiatives in the area (including those displayed in records of Web archive collections, such as Cornell University and Congress libraries), which were described in the paper. In a non-exhaustive way, Table 1 summarizes the basic aspects of the aforementioned standards and the metadata elements (or the semantic units for PREMIS) considered, for this research, as important in the preservation of Web content.

Table 1. Patterns and metadata elements to support digital preservation in web archiving. (continues)

Standard	Characteristics	Metadata elements useful for digital preservation in web archiving	
Qualified DC (version 1.1)	- XML Schema or other syntax considered flexible, extensible, simple and interoperable; and - Applicable to discovery of web resources and for web archiving for their general use by Archive-It users.	▪ Title ▪ Creator ▪ Subject) ▪ Description ▪ Contributor ▪ Date ▪ Type ▪ Format	▪ Identifier ▪ Source ▪ Language ▪ Relation ▪ Coverage ▪ Rights ▪ RightsHolder ▪ Provenance

4 Using PREMIS with METS. Available at: https://www.loc.gov/standards/premis/premis-mets.html. Access on: 23 July 2021.

5 Registered profiles. Available at: https://www.loc.gov/standards/mets/mets-registered-profiles.html. Access on: 23 July 2021.

MODS (version 3.7)	- XML schema derived from MARC 21 seen as richer than DC and simpler than full MARC; - Can be used together with MADS and as a METS extension scheme; and - Applicable to digital library objects and to archived sites, such as those in the Library of Congress Web Archive collections.	▪ Title information (<titleInfo>) ▪ Name (<name>) ▪ Resource type (<typeOfResource>) ▪ Genre (<genre>) ▪ Origin Information (<originInfo>) ▪ Language (<language>) ▪ Physical Description (<physicalDescription>) ▪ Abstract (<abstract>) ▪ Table of Contents (<tableOfContents>) ▪ Note (<note>) ▪ Subject (<subject>)	▪ Related item (<relatedItem>) ▪ Identifier (<identifier>) ▪ Location (<location>) ▪ Access Condition (<accessCondition>) ▪ Part (<part>) ▪ Extension (<extension>) ▪ Record Information (<recordInfo>)
EAD3 (version 1.1.1)	- Thorough XML schema and DTD, which is compliant with ISAD(G) archival description standard; - Can be used in conjunction with EAC-CPF and includes both indication of corresponding elements in MARC, MODS, ISAD(G) and HTML as mappings (crosswalks) for MARC 21, MODS and ISAD(G) ; and - Applicable to the coding of archival search instruments, such as the California Online Archive Search Instruments that provide detailed descriptions of the collections from the University of California Irvine web archive in the United States.	▪ Unit title (<unittitle>) ▪ Origin (<origination>) ▪ Personal Name (<persname>) ▪ Organization Name (<corpname>) ▪ Family Name (<famname>) ▪ Controlled Access Headers (<controlaccess>) ▪ Abstract (<abstract>) ▪ Accruals (<accruals>) ▪ Acquistiotion Information (<acqinfo>) ▪ Biography or history (<bioghist>) ▪ Scope and Contents (<scopecontent>) ▪ Custodian History (<custodhist>) ▪ Descriptive Identification Note (<didnote> ▪ Other Descriptive Data (<odd>) ▪ Unit Date (<unitdate>)	▪ Genre/Physical Characteristic (<genreform>) ▪ Physical Description (<physdesc>) ▪ Digital Archive Object (<dao>) ▪ Unit Identification (<unitid>) ▪ Material Language (<langmaterial>) ▪ Language (<language>) ▪ Physical Structure (<physdescstructured>) ▪ Related Material (<relatedmaterial>) ▪ Access Restrictions (<accessrestrict>) ▪ Use Restrictions (<userestrict>) ▪ Physical Location (<physloc>) ▪ Process Information (<processinfo>) ▪ Repository (<repository>) ▪ Maintenance Agency (<maintenanceagency>) ▪ Maintenance History (<maintenancehistory>)

Table 1. Patterns and metadata elements to support digital preservation in web archiving. (conclusion)

Standard	Characteristics	Metadata elements useful for digital preservation in web archiving	
VRA *Core* (version 4.0)	- Simple XML schema that can be used together with the CCO, with the indication of equivalent elements in the CCO, CDWA and DC; and - Applicable to the description of original cultural works and their reproductions, such as certain Cornell University library collections, which	▪ Work, collection or Image (<work>, <collection>, <image>) ▪ Agent (<agent>) ▪ Cultural Context (<culturalContext>) ▪ Date (<date>) ▪ Description (<description>) ▪ Inscription (<inscription>) ▪ Localization (<location>) ▪ Material (<material>) ▪ Measurements (<measurements>)	▪ Relation (<relation>) ▪ Rights (<rights>) ▪ Source (<source>) ▪ State/Edition (<stateEdition>) ▪ Period/Style (<stylePeriod>) ▪ Subject (<subject>) ▪ Techniquue (<technique>) ▪ Text Reference (<textref>) ▪ Title (<title>) ▪ Work Type (<worktype>)

	have a web archive of their websites.		
PREMIS (version 3.0)	- XML schema that focuses on the preservation repository and its management; - Can join other standards such as MODS, DC, EAD, METS etc. to cover metadata outside its scope and additional functions; and - Applicable to support the preservation of digital objects, such as in the description of rendering environments for web content.	▪ objectIdentifier/Category ▪ preservationLevel ▪ significantProperties ▪ objectCharacteristics ▪ originalName ▪ storage ▪ signatureInformation ▪ environmentFunction/Designation/Registry/Extension ▪ relationship ▪ linkingEventIdentifier/RightsStatementIdentifier	▪ eventIdentifier/Type/DateTime ▪ eventDetailInformation/OutcomeInformation ▪ linkingAgentIdentifier/ObjectIdentifier ▪ agentIdentifier/Name/Type/Version/Note/Extension ▪ linkingEventIdentifier/RightsStatementIdentifier/EnvironmentIdentifier ▪ rightsStatement/Extension
METS (version 1.12.1)	- Flexible XML schema that organizes and links metadata forms to objects in a system; - Can structure the OAIS PSI, PAI or PDI packages and include standards in the Descriptive Metadata section, such as DC, and have PREMIS in the Administrative Metadata section; and - Applicable to the transfer and digital preservation of web capture content (websites) between repositories through the METS profile of the ECHO DEPository project.	▪ Agent (<agent>) ▪ Alternative Identifier (<altRecordID>) ▪ Metadata Reference (<mdRef>) ▪ Metadata Wrapper (<mdWrap>) ▪ Technical Metadata (<techMD>) ▪ Intellectual Property Rights Metadata (<rightsMD>) ▪ Source Metadata (<sourceMD>) ▪ Digital Provenience Metadata (<digiprovMD>) ▪ File Group (<fileGrp>) ▪ File (<file>)	▪ File Localization (<FLocat>) ▪ File Contents (<FContent>) ▪ Component Byte Stream (<stream>) ▪ Transforming File (<transformFile>) ▪ Division ( ) ▪ File Pointer (<fptr>) ▪ METS Pointer (<mptr>) ▪ Structural Mapping Link (<smLink>) ▪ Behavior (<behavior>) ▪ Interface Definition (<interfaceDef>) ▪ Mechanism (<mechanism>)

Source: The authors.

First, it is important to note that all the metadata patterns analyzed in Table 1 are expressed in XML syntax and, to some extent, are flexible and/or extensible. As an open and readable standard for computers and humans, XML meets the needs of digital preservation and allows the description and exchange of diverse types of data on the Web and in other environments; furthermore, it facilitates the integration and combined use of various schemas based on this same language, such as external schemas for joint use with METS, which include EAC-CPF, MARC, MIX, etc. By the way, there is the Dublin Core Metadata Initiative – DCMI,

in the case of DC, and the Network Development and MARC Standards Office of the American Library of Congress, for the other standards analyzed (except the EAD and VRA Core, which are maintained, in that order, by the Technical Subcommittee for Encoded Archival Standards – TS-EAS of the Society of American Archivists – SAA and by the VRA Core Oversight Committee), which standardize the description and representation of information through value encoding schemes.

Secondly, the joint use of various metadata standards, driven by external metadata indications, by equivalent/corresponding elements and by mappings (crosswalks) or by the common adoption of syntaxes, norms and vocabularies, which portray the flexibility and extensibility of schemas and increase data interoperability is acceptable, given the high complexity of the types of resources to be described and the different stages of long-term digital preservation and Web archiving processes of preservation, we further infer that all metadata classes – descriptive, markup languages, etc. – are vital to the achievement of Web content preservation. For now, it is likely that we cannot establish what the only standard that fully guarantees digital preservation is, but existing standards can be completed to document the information required in preservation management and the usable access of complex digital objects such as websites.

Among the assiduous metadata in the analyzed patterns are the identifiers, which can be contained in the objectIdentifier PREMIS unit and in the Identifier and DC Relation elements; Related Item, MODS Identifier and Location; Related Material, Unit Identification and Digital Archive Object EAD; Textual Reference VRA; File Location, Metadata Reference, Interface Definition, Mechanism and METS Indicator. Judging the relationships of a single website being described with any collection to which it belongs or with other resources, identifiers provide the unique and distinctive identification of the resource to which the metadata refers, such as its electronic location. Therefore, the record of an archived website URL (access, capture etc.) and the URL for the related resource reflect the metadata defined in the Preservation Description Information (reference and context) of the OAIS information model and preservation principles to maintain context and to identify and locate objects.

In fact, DC, MODS, EAD and VRA Core are more apt to describe digital resources for the purposes of discovery, retrieval, presentation and interoperability. Even though the scopes of these metadata standards are inherently aimed at the access stage rather than just long-term preservation, some of their descriptive elements are useful in supporting PREMIS preservation metadata. Thus, the information provided by them permeates aspects of representation and preservation, such as technical characteristics and dependencies, changes made, chain of custody or ownership, origin, physical and logical structural relations, rights, etc., which are relevant in the management of objects archived digital data and which, to some degree, translate part of the contours of OAIS preservation metadata and the principles of digital preservation to ensure the reliability, authenticity and integrity of objects and to maintain their context, provenance and retrieval to the over time.

In light of digital preservation in Web archiving and based on the examples of archived Web content descriptions by Dappert *et al*. (2013), Digital Library Federation (2010), Dooley and Bowers (c2018), Habing (2006) and Library of Congress (2018), we distributed the metadata elements indicated in Table 1 of the DC, MODS, EAD and VRA Core standards according to the information they can record for archived websites and site collections:

- Title, Creator, Subject, Contributor and Language DC; Title, Location, Name, Language, Index, Subject and Part MODS information; Unit Title, Origin, Digital Archive Object, Personal Name, Organization Name, Family Name, Controlled Access Headings, Material Language, Language and EAD Repository; and Period/Style, Agent, Location, Subject and VRA Title, which express the name given to the described resource, the topic, the language, the person or organization responsible for creating its intellectual content or making contributions to it and the institution or repository that holds the resource. For example, the name and language

of the site or archived collection, the entity that created its content or made secondary contributions and the institution responsible for its selection, curation or management, in addition to thematic subjects, names of geographic places and entities used to main topic describing the archived content or site.

- Description, Date and Coverage DC; Source Information, Summary and Note MODS; Unit Date, Descriptive Identification Note, Other Descriptive Data, Process Information, Summary, Additions, Biography or History, Scope and Content, Custody History and Acquisition Information EAD; and Cultural Context, Location, Description, Date and Period/VRA Style, which provide an account of the content, scope and context of the described resource, a period of time linked to an event in the life cycle of the resource and the chain of custody , the origin and the topic or spatiotemporal scope of the resource. For example, the indication of copyright dates or when the site was started/inactivated, was/began to be archived and the URL was captured (with its frequency), as well as provenance (if a site is part of a collection broader theme, etc.) and a legal decree or other reason for selecting the site or content for archiving.
- DC Source and Provenance; MODS Source Information; Date of the EAD Unit; and VRA Source, which expresses a reference to the source of the recorded information about the described resource and about another resource from which it is derived, changes in custody and ownership of the resource since its creation; and the origin of the resource, including place of origin/publication, publisher and associated dates, such as the date and place the archived site was created/issued and the date of its capture.
- DC Relation; Related Item MODS; EAD Related Material; and VRA Relation, which provide a reference to another resource related to the resource being described. For example, the whole/part relationships between a single archived site and any collection of archived sites it belongs to (with its title included), between an archived site and a collection of analog archives or other digital materials such as web pages constituents of the website and the images and videos that make up the website.
- Rights and DC Rights holder; MODS Access Condition; EAD Access Conditions and Conditions of Use; and VRA Rights, which record the rights to the resource described, the person/organization that has or administers those rights, the restrictions (or lack thereof), and the conditions that affect access, rendering, and use of the resource. For instance, the indication for use on site and a period in which the archived content or website is restricted, if access to the content is open and if the rights holders allow reuse after access.
- DC Type and Format; Resource Type, Genre and Physical Description MODS; Controlled Access Headers, Gender/Physical Characteristic, Structured Physical and Physical Description EAD; and Work, Collection or Image, Material, Measures, Type of Work and VRA Technique, which expose the nature of the described resource and its format, dimensions, technique and style. For example, indicating whether the archived content is website, web archives, social media, etc. and that it is a collection with a particular number of archived sites.
- Extension and MODS Registration Information; Physical Location, Maintenance Agency and EAD Maintenance History; and VRA Registration and Status/Edition, which document information about the resource using more than one scheme, the physical location of the resource and the institution/service responsible for its creation, maintenance and dissemination, the identification of the edition of the resource and the its history of creation, revisions, updates and other changes, as well as information for the management and interpretation of the metadata record, such as, for example, the origin of the website record or the archived collection (machine-generated or not, etc.) and its language, date it was first created, organization that created or changed its

original version, and rules used for the content of the description (i.e., controlled vocabularies, cataloging standards, etc.).

Despite being out of their purview, technical metadata MIX, Technical Metadata for Text (TextMD), Multimedia Content Description Interface (MPEG-7), Audio/Video Technical Metadata Extension Schema (Audio/VideoMD) and MADS and authority data EAC-CPF can also be used with PREMIS together with METS to record the circumstances of creation (date of creation, name of the creation device, etc.), the history of changes made (documented, authorized, etc.), the characteristics and technical dependencies (size, hardware, etc.) and other aspects of audiovisual, text and other formats integrated in the archived sites, as well as the recording of data on agents with roles in creation and contribution, selection, curatorship or management , in the rights, in the rendering and in the actions that affect these materials. Therefore, these standards support the interoperability, management and preservation of complex digital objects, such as websites that include various formats and types of content, and should consider them for the definition and validation of the origin, authenticity and integrity of their contents.

In turn, PREMIS portrays the practical use of the preservation metadata concepts outlined in the OAIS information model and, subsequently, reflects the requirements and principles of digital preservation, which makes all its semantic units important for the long-term preservation of archived websites. Therefore, drawing on Guenther, Dappert and Peyrard (c2016), who illustrate relationships between the semantic units of the PREMIS data dictionary and the OAIS information categories, we highlight the significantProperties, environmentFunction/Designation/Registry/Extension and relationship units (context and provenance information, structural information and other OAIS representations) that may detail, for example, that only the content needs to be maintained for a webpage, containing animations that were not taken as vital; the environment that supports the rendering and running of a website; and the relationships involving technical environment and structural relationships between integral parts of a website.

Finally, in an OAIS repository, METS serves as a central schema for managing archived websites and transferring these objects between systems (or between systems and their users), including DC, MODS, EAD, VRA Core, MARC XML, etc. in the Descriptive Metadata section such as PREMIS, MIX, TextMD, AudioMD and VideoMD etc. in the Administrative Metadata section of the METS document. In the second section, the PREMIS units (for example, eventIdentifier/Type/DateTime or eventDetailInformation/OutcomeInformation) record, in the Digital Provenance Metadata (<digiprovMD>) element, any preservation-related actions performed on the various files that make up a site or which modifications were made to a digital object (website) and/or its constituent parts during its lifecycle which, according to Digital Library Federation (2010), can be used to judge how these processes have altered or corrupted the object's ability to represent with accuracy the original item.

Thus, descriptive and administrative metadata (<mdRef> and <mdWrap>; <techMD>, <rightsMD>, <sourceMD> and <digiprovMD>) can be external to the METS document, with the latter recording original/derivative relationships between files, how files were created and stored, etc. Useful for digital preservation requirements, the METS Header and File sections (<agent> and <altRecordID>; <transformFile>, <fileGrp>, <file>, <FLocat>, <FContent> and <stream>) include metadata about the METS document itself and list (by format etc.) files that make up the content of websites. Other sections are Behavior (<behavior>, <interfaceDef> and <mechanism>) for rendering or displaying the site and Structural Map and Structural Links (, <fptr> and <mptr>; <smLink>), which order the hyperlinks between files that make up the objects or between other objects, such as a Web page with an image linked to another Web page, recording the hypertext structure of the archived sites separate from the HTML files of the site itself and which can be shown to users for their understanding and for content navigation (DIGITAL LIBRARY FEDERATION, 2010; LIBRARY OF CONGRESS, 2017).

6 FINAL CONSIDERATIONS

In practical terms, the main problems of digital preservation derive from the particularities of the objects to which it is intended to maintain access, recovery and use over time. One of the examples of complex digital objects are websites that contain both a wide range of hypertext links to allow navigation from one web page to another, as well as various files and formats with a high dependence on technologies for their access, interpretation, rendering and use, which become obsolete over time; in fact, they are subject to the dynamics and ephemerality of the Web, where their contents are created and published and, therefore, they are lost or quickly undergo changes in their original form. Therefore, these facets force us to reflect on the issues of authenticity, integrity and context of archived websites, and also to elucidate the distinctions between live websites and their archived fixed versions, such as the usefulness of mixed description approaches for a single site or an archived collection due to its heterogeneity.

As one of the aspects of ensuring digital preservation, the adoption of metadata for long-term preservation helps in decision-making and in the control of legal requirements, versions, continuity of access, use and interpretation, and other issues related to archiving of objects in systems. Metadata schemas can provide interoperability of objects between repositories/services, encourage the common use of vocabularies, thesaurus and controlled lists (data value standards) – such as LCSH, Internet MIME types, ISO 639. –, or standards, rules and bibliographic and archival cataloging codes (data content standards) – such as RDA – and allow joint description or inclusion of metadata from other XML schemas with pointers to external metadata, as in the MODS Extension element.

In summary, the research carried out identified, systematized and analyzed patterns and metadata schemas for Web archiving, debated in Information Science and in related areas. In addition, it indicated that the descriptive and technical metadata DC, MODS, EAD, VRA Core, MIX etc. and the MADS and EAC-CPF authority data have an application more focused on supporting PREMIS and METS, either in allowing identification and location or in providing technical data, rendering, integrity and fixity, rights and agents with functions in the actions that affect archived sites. It also concluded that, by incorporating descriptive, structural and administrative (and preservation) metadata, such as PREMIS, METS is useful in simplifying the ordering and management of the constituent parts of sites and their metadata, hierarchically linking the different files (texts, images etc.) that make up the sites and, in addition, manage such complex objects, acting as a PSI, PAI and PDI in an OAIS.

On the other hand, through the referenced literature, we find some disadvantages of the metadata standards identified in this work that can be adapted and/or applied to digital preservation and Web archiving: DC suffers criticisms to its structure and to the very simplistic and generic set of elements (especially, compared to other formats, such as MARC); in MODS, conversions of original MARC records to this standard and then back to MARC can result in loss of data or some loss of specificity in the markup; in ODL, the absence of resources and knowledge in an institution can influence its use; VRA Core has specificity, imposes certain restrictions on linking to non-VRA Core records, and is less common compared to other formats; in PREMIS, the lack of training/expertise and integration with the existing system can bring barriers to its adoption; and METS has a flexibility that causes interoperability issues and also an imperfect correlation with PREMIS, including duplications between the elements of these two metadata schemas.

Still, as a required research on the issue of metadata within Web archiving, Dooley and Bowers (c2018) cite, for example, the undefined boundaries between descriptive metadata and other metadata categories, such as tracking dates, which are clearly both descriptive and technical; and what types of metadata to capture, including how they are extracted, merged with descriptive metadata, and made intelligible to end users. Reinforcing the authors' argument, we

propose, in addition to further studies on this recent topic in national and international scientific literature, that new researches explain the dilemmas and solutions taken in the implementation of each standard for the scope of archived content on the Web. Research should examine how the metadata of the standards identified in the work can be better harmonized, avoiding problems of duplication and redundancies, as the analysis of the results indicated that DC, MODS, EAD and VRA Core supported METS and PREMIS in the discovery and in documenting technical aspects of archived websites and in proving their authenticity, context and provenance.

Anyway, different types of metadata are important in web archiving, but this work focused on descriptive and administrative (mainly preservation) metadata. Certain metadata elements or semantic units of the identified patterns could be flagged in this research as being useful for the preservation of websites in digital archiving systems. For example, in the DC, the elements indicated in Table 1 include information defined in the PREMIS data dictionary units, such as copyright and its holders, the unique and persistent identification, the whole/part and derivation relationships, and the technical dependencies of the digital object. By the way, DC proved to be an exponent for Web archiving due to its similarities with the WAN elements of Dooley and Bowers (c2018) and, in the use of elements by Kim and Lee (2007), in the Internet Archive, in Arquivo.pt and in other notable initiatives in the area.

Thus, the results of the work provide a theoretical, technical and structured support of metadata standards and schemas, which can be used in Web archives designed to serve the preservation and provide lasting access to archived Web contents. Both the metadata elements and the semantic units pointed out in the research for digital preservation in Web archiving will collaborate to the choice of metadata standards according to the needs of public, private, non-profit, research and cultural heritage organizations that are interested and/or involved in national and international initiatives in the area or, still, for the perception of the information to be foreseen and required to ensure the description, preservation and consistent management of the archived sites in a system that were selected and collected from an electronic domain, event, location or topic (science and technology etc.).

Therefore, it is evident that the guarantee of digital preservation in Web archiving will only be feasible with the effective adoption of metadata standards in support of archiving administration and maintenance of permanent and usable access to Web contents over time. These description structures will define the identity and persistence, coherence and understandability, access and representation, functionalities, authenticity, integrity and reliability, context and provenance of selected websites, collected and stored in information systems, information for preservation in addition to determining the discovery, retrieval, presentation, navigation and archivability of websites, such as semantic interoperability between systems.

REFERENCES

ALEMNEH, Daniel Gelaw; HASTINGS, Samantha Kelly. Exploration of adoption of preservation metadata in cultural heritage institutions: case of PREMIS. **Proceedings of the American Society for Information Science and Technology**, v. 47, n. 1, p. 1-8, Nov./Dec. 2010.

ALLISON-BUNNELL, Jodi. Review of Encoded Archival Description Tag Library: version EAD3. **Journal of Western Archives**, v. 7, n. 1, p. 1-4, 2016.

ALVES, Rachel Cristina Vesú. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2010.

ALVES, Rachel Cristina Vesú. Metadados para representação e recuperação da informação em ambiente Web. *In*: MARINGELLI, Isabel Cristina Ayres da Silva. (org.). **IV Seminário Serviços de Informação em Museus**: informação digital como patrimônio cultural. São Paulo: Pinacoteca de São Paulo, 2017. p. 95-106.

ARQUIVO.PT. **Metadados acerca dos conteúdos**. [*S. l.*], ago. 2018.

BAILEY, Jefferson; LACALLE, Maria. Don't warc away: preservation metadata and web archives. *In*: AMERICAN LIBRARY ASSOCIATION (ALA) ANNUAL CONFERENCE, 16., June 2015, San Francisco, California. **Proceedings** […]. San Francisco, California: ALA, 2015. p. 1-46.

BANOS, Vangelis *et al*. CLEAR: a credible method to evaluate website archivability. *In*: INTERNATIONAL CONFERENCE ON PRESERVATION OF DIGITAL OBJECTS (iPRES), 10, May c2013, Lisboa, Portugal. **Proceedings** […]. Lisboa, Portugal: iPRES, 2010. p. 9-18.

CANTARA, Linda. METS: the metadata encoding and transmission standard. **Cataloging & Classification Quarterly**, Philadelphia, v. 40, n. 3/4, p. 237-253, 2005.

CAPLAN, Priscilla. **Understanding PREMIS**. [Washington, DC]: Library of Congress Network Development and MARC Standards Office, 2017. 22 p.

CASTRO, Fabiano Ferreira de Castro. **Elementos de interoperabilidade na catalogação descritiva**: configurações contemporâneas para a modelagem de ambientes informacionais digitais. 2012. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2012.

CHAN, Lois Mai; ZENG, Marcia Lei. Metadata interoperability and standardization: a study of methodology part i: achieving interoperability at the schema level. **D-Lib Magazine**, [*S. l.*], v. 12, n. 6, June c2006.

CHEN, Mingyu; REILLY, Michele. Implementing METS, MIX, and DC for sustaining digital preservation at the University of Houston Libraries. **Journal of Library Metadata**, [*S. l.*], v. 11, n. 2, p. 83-99, May 2011.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. The evolution of web archiving. **International Journal on Digital Libraries**, v. 18, n. 3, p. 191-205, Sept. 2017.

DAPPERT, Angela *et al*. Describing and preserving digital object environments. **New Review of Information Networking**, Philadelphia, v. 18, n. 2, p. 106-173, Oct. 2013.

DAPPERT, Angela; ENDERS, Markus. Digital preservation metadata standards. **Information Standards Quarterly (ISQ)**, v. 22, n. 2, p. 4-13, spring 2010.

DI PRETORO, Emmanuel; GEERAERT; Friedel. Behind the scenes of web archiving: metadata of harvested websites. Archives et Bibliothèques de Belgique – Archief – En Bibliotheekwezen in Belgie; Archief, in press, trust an Undertanding: The value of metadata en a digitally joined-up world. 2019.

DIGITAL LIBRARY FEDERATION. **<METS> Metadata Encoding and Transmission Standard**: primer and reference manual. Version 1.6. [Washington, DC], 2010. 144 p.

DIGITAL PRESERVATION COALITION. **Metadata**. [Glasgow, Scotland], [201-?]. 2 p. (Digital Preservation Topical Notes, 5).

DOOLEY, Jackie M. *et al*. Developing web archiving metadata best practices to meet user needs. **Journal of Western Archives**, [Provo], v. 8, n. 2, p. 1-14, 2017.

DOOLEY, Jackie; BOWERS, Kate. **Descriptive metadata for web archiving**: recommendations of the oclc research library partnership web archiving metadata working group. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 53 p.

DOORN, Peter; TJALSMA, Heiko. Introduction: archiving research data. **Arch Sci**, v. 7, p. 1-20, Sept. 2007.

DUBLIN CORE METADATA INITIATIVE. About DCMI. **DCMI History**. [*S. l.*], June c2020a.

DUBLIN CORE METADATA INITIATIVE. DCMI Usage Board. Specifications. **DCMI Metadata Terms**. [*S. l.*], Jan. 2020b.

DUBLIN CORE METADATA INITIATIVE. DCMI Usage Board. Specifications. **Dublin Core Metadata Element Set, Version 1.1***:* reference description. [*S. l.*], June 2012.

EIDSON, Jennifer G.; ZAMON, Christina J. EAD twenty years later: a retrospective of adoption in the early twenty-first century and the future of ead. **The American Archivist**, v. 82, n. 2, p. 303-330, 2019.

EÍTO-BRUN, Ricardo. A metadata infrastructure for a repository of civil engineering records: eac-cpf as a cornerstone for content publishing. **Journal of Archival Organization**, v. 12, n. 1-2, p. 62-76, 2015.

FORMENTON, Danilo *et al*. Os padrões de metadados como recursos tecnológicos para a garantia da preservação digital. **Biblios**, Pittsburgh, n. 68, p. 82-95, jul. 2017.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010. 184 p.

GILLILAND, Anne J. Setting the stage. *In*: BACA, Murtha. (ed.). **Introduction to metadata**. 3rd ed. Los Angeles, California: Getty Publications, c2016. 92 p.

GUENTHER, Rebecca S. MODS: the metadata object description schema. **Portal**: Libraries and the Academy, v. 3, n. 1, p. 137-150, Jan. 2003.

GUENTHER, Rebecca Squire; DAPPERT, Angela; PEYRARD, Sébastien. An introduction to the PREMIS data dictionary for digital preservation metadata. *In*: GUENTHER, Rebecca Squire; DAPPERT, Angela; PEYRARD, Sébastien. **Digital preservation metadata for practitioners**. Cham, Switzerland: Springer, Dec. c2016. p. 23-36.

GUENTHER, Rebecca; MYRICK, Leslie. Archiving web sites for preservation and access: MODS, METS and MINERVA. **Journal of Archival Organization**, v. 4, n. 1/2, p. 141-166, 2007.

HABING, Thomas G. **ECHO Dep METS Profile for Web Site Captures**. [*S. l.*], 2006.

HARPER, Corey A. Dublin Core Metadata Initiative: beyond the element set. **Information Standards Quarterly (ISQ)**, v. 22, n. 1, p. 19-28, winter 2010. Acesso em: 2 jun. 2020.

INTERNET ARCHIVE. Internet Archive APIs. About Archive.org metadata. **Internet archive metadata**. [San Francisco, California], Dec. 2018.

KIM, Heejung; LEE, Hyewon. Development of metadata elements for intensive web archiving. **Journal of the Korean Society for Information Management**, Songdo, South Korea, v. 24, n. 2, p. 143-160, June 2007.

LAVOIE, Brian; GARTNER, Richard. Preservation metadata. 2nd edition. **DPC Technology Watch Report**, v. 13, n. 3, p. 1-36, May c2013.

LIBRARY OF CONGRESS. **Development of the Encoded Archival Description DTD**. Dec. 2013.

LIBRARY OF CONGRESS. **METS**: an overview & tutorial. [Washington, DC], Mar. 2017.

LIBRARY OF CONGRESS. **MODS user guidelines**. MODS elements and attributes. Version 3. [Washington, DC], Aug. 2018.

LIBRARY OF CONGRESS. **MODS**: uses and features. [Washington, DC], Feb. 2016.

LIBRARY OF CONGRESS. Programs. Web archiving. About this program. Web archives. **Collections with web archives**. [Washington, DC], [2021].

LIMA, Fábio Rogério Batista; SANTOS, Plácida Leopoldina V. A. C.; SANTARÉM SEGUNDO, José Eduardo. Padrão de metadados no domínio museológico. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 21, n. 3, p. 50-69, jul./set. 2016.

LUBAS, Rebecca L.; JACKSON, Amy S.; SCHNEIDER, Ingrid. Using VRA Core 4.0. *In*: LUBAS, Rebecca L.; JACKSON, Amy S.; SCHNEIDER, Ingrid. **The metadata manual**: a practical workbook. Oxford, UK: Chandos Publishing, 2013. p. 135-164.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 8. ed. atual. São Paulo: Atlas, 2017. 368 p.

MÁRDERO ARELLANO, Miguel Ángel. **Critérios para a preservação digital da informação científica**. 2008. Tese (Doutorado em Ciência da Informação) - Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2008.

MASANÈS, Julien. **Web Archiving**. Berlin: Springer, c2006. 234 p.

MCCALLUM, Sally H. An introduction to the metadata object description schema (MODS). **Library Hi Tech**, v. 22, n. 1, p. 82-88, 2004.

MCDONOUGH, Jerome P. METS: standardized encoding for digital library objects. **International Journal on Digital Libraries**, v. 6, n. 2, p. 148-158, April 2006.

MELO, Jonas Ferrigolo; ROCKEMBACH, Moisés. Arquivabilidade de websites para preservação digital: estudo a partir da área da saúde. **Reciis** – Rev Eletron Comun Inf Inov Saúde, v. 14, n. 3, p. 529-545, jul./set. 2020.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **Understanding metadata**. Bethesda, Maryland: NISO Press, c2004. 16 p.

NATIONAL LIBRARY OF NEW ZEALAND. **Metadata standards framework**: preservation metadata (revised). Wellington, New Zealand: National Library of New Zealand, June 2003. 50 p.

PALA, Francesca. Lo standard EAD3 per la codifica dei dati archivistici: qualche novità e molte conferme. **JLIS.it**, Macerata, v. 8, n. 3, p. 148-176, Sept. 2017.

PENNOCK, Maureen. Web-Archiving. **DPC Technology Watch Report**, v. 13, n. 1, p. 1-45, Mar. c2013.

PREMIS EDITORIAL COMMITTEE. **PREMIS data dictionary for preservation metadata**. Version 3.0. [*S. l.*: *s. n.*], Nov. 2015. 273 p.

RILEY, Jenn. **Understanding metadata**: what is metadata, and what is it for? Baltimore, Maryland: National Information Standards Organization (NISO), c2017. 45 p.

ROCKEMBACH, Moises; PAVÃO, Caterina Marta Groposo. Políticas e tecnologias de preservação digital no arquivamento da web. **RICI**: R.Ibero-amer. Ci. Inf., Brasília, v. 11, n. 1, p. 168-182, jan./abr. 2018.

ROWELL, Chelcie Juliet; KREWER, Drew. Preservation metadata for complex digital objects. A Report of the ALCTS PARS Preservation Metadata Interest Group Meeting. American Library Association Annual Conference, San Francisco, June 2015. **Technical Services Quarterly**, v. 33, n. 2, p. 179-183, Mar. 2016.

SAMOUELIAN, Mary; DOOLEY, Jackie. **Descriptive metadata for web archiving**: review of harvesting tools. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 23 p.

SAYÃO, Luís Fernando. Uma outra face dos metadados: informações para a gestão da preservação digital. **Enc. Bibli**: R. Eletr. Bibliotecon. Ci. Inf., Florianópolis, v. 15, n. 30, p. 1-31, 2010.

SENANDER III, Mathew. Converting vra core records to marc records: a study in crosswalking. **Library Philosophy and Practice**, Lincoln, Dec. 2013.

SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 24. ed. rev. e atual. São Paulo: Cortez, 2016. 320 p.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. rev. e atual. Florianópolis: Universidade Federal de Santa Catarina (UFSC), 2005. 139 p.

SOCIETY OF AMERICAN ARCHIVISTS. Technical Subcommittee for Encoded Archival Standards. **Encoded Archival Description Tag Library**: version EAD3 1.1.1. Chicago, Dec. 2019. 422 p.

TRUMAN, Gail. **Web archiving environmental scan**. Harvard Library Report. [Cambridge, Massachusetts]: Harvard University, Jan. 2016. 83 p.

VEIKKOLAINEN, Petteri; LAGER, Lassi. Long-term preservation of the finnish web archive. *In*: INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC) GENERAL ASSEMBLY, 10., April 2016, Reykjavik, Iceland. **Proceedings** […]. Reykjavik, Iceland: IIPC, 2016. p. 195-203.

VELLUCCI, Sherry L. Metadata and authority control. **Library Resources & Technical Services (LRTS)**, [Chicago], v. 44, n. 1, p. 33-43, Jan. 2000.

VENLET, Jessica *et al*. **Descriptive metadata for web archiving**: literature review of user needs. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 48 p.

VISUAL RESOURCES ASSOCIATION. **An introduction to VRA Core**. VRA Core 4.0 introduction. [*S. l.*], Oct. 2014. 2 p.

VISUAL RESOURCES ASSOCIATION. **VRA Core 4.0 element description**. [*S. l.*], May 2007. 37 p.

ZENG, Marcia Lei.; QIN, Jian. **Metadata**. New York, United States: Neal-Schuman Publishers, June 2008. 365 p.