

Metadata Principles, Guidelines and Best Practices: A Case Study of Brazil and Sri Lanka

Francisco Carlos Paletta

<https://orcid.org/0000-0002-4112-5198>

University of São Paulo, Brazil

fcpaletta@usp.br

Chiranthi Wijesundara

<https://orcid.org/0000-0001-9254-5750>

University of Colombo, Sri Lanka

chiranthi@lib.cmb.ac.lk

Abstract

The concepts, guidelines, and best practices related to metadata are well-known factors in the global context. Hence, the weight given to this domain is varied in each country. The study investigates some fundamental conceptions of metadata and related standards through various literature and presents some best practices of metadata related to Brazil and Sri Lanka. Both countries have some initiatives related to library and geospatial data domains. The scale of these projects may be different, but some similarities are visible in both scenarios. Libraries use various metadata standards to organize and retrieve resources and this applies to both countries. Compared to Brazil, Sri Lankan Library, Archive and Museum (LAM) awareness of metadata is confined to MARC standards. Similar to Brazil, many institutions in Sri Lanka are maintaining Dspace repositories that use a qualified Dublin Core-based metadata schema. Some professionals in the information science sector are aware of Dublin Core standards, but the use cases are very rare. Based on the above best practices, awareness of Dublin Core metadata standards in Brazil is wider compared to Sri Lanka. However, this scenario should be further investigated thoroughly. Finally, awareness of basic conceptions and standards related to metadata is a key factor when it comes to conducting more research in the domain.

Keywords: Best Practices, Metadata, Brazil, Sri Lanka, Standards

1. Introduction

During the past few decades, metadata concepts and guidelines have changed rapidly due to the development of subject and technology domains. The urgency of this rapid change is diverse from one country to another. Technological, socioeconomic, external, and internal factors have affected this diversity. Nevertheless, the core concepts and principles of metadata are pivotal in the development of projects related to metadata within the rapidly changing information environment.

“Metadata” is commonly understood as “data about data,” contributing valuable information about the context in which the data was collected. For instance: information about the source of the data, the date it was collected, people involved, and the methodology used to collect are some metadata that can be gathered to make data contextual. This can help analysts better understand the data, and its quality, and be able to make decisions safely.

Creating descriptive metadata serves a crucial role in different aspects as follows (NISO, 2004).

- i. Resource Discovery: Metadata functions similar to effective cataloging, facilitating resource discovery by allowing resources to be located based on relevant criteria, identifying them, grouping similar resources, distinguishing dissimilar ones, and providing location information. This is extremely useful in the Library and archival domains.
- ii. Organizing Electronic Resources: With the exponential growth of web-based resources, aggregate sites or portals become increasingly valuable in organizing links based on audience or topic. While static web pages can list resources with hardcoded names and locations, dynamically building these pages from metadata stored in databases is more efficient and common.
- iii. Interoperability: Describing resources with metadata enables understanding by both humans and machines, fostering interoperability. This facilitates data exchange across multiple systems with different hardware, software platforms, data structures, and interfaces with minimal loss of content and functionality.
- iv. Digital Identification: Most metadata schemes include elements such as standard numbers to uniquely identify the work or object. These persistent identifiers (e.g., DOIs/ PURLs) ensure the accessibility of resources even if their locations change.
- v. Archiving and Preservation: Metadata plays a crucial role in archiving and preservation efforts by tracking the lineage of digital objects, detailing their physical characteristics, and documenting their behavior for emulation on future technologies.
- vi. Support AI: Metadata provides critical context and information about the data being used to train and operate AI algorithms. It is the framework that shapes how AI systems process, interpret and generate insights from available data in order to respond to queries.

Varying from libraries, archives, museums, publishing, broadcasting, Data warehousing, healthcare industry, environmental agencies, government organizations and various other organizations that basically work with data/ information utilize diverse metadata in various quantities.

Based on the usage and the requirements of metadata various concepts, standards, models and tools have been developed. For instance, a metadata standard is a requirement which is intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users.

This article investigates some common concepts and guidelines/standards related to metadata in different domains. Finally, the article presents two different scenarios of metadata best practices in Brazil and Sri Lanka.

2. Metadata Conceptions and Standards

2.1. Definitions

Although now widely recognized, the term *metadata* was initially coined by Jack E. Myers in 1969. Later in 1986, METADATA® was registered as a US trademark (Greenberg, 2005). Metadata emerged as a prominent concept in the 1980s and 90s and has been adopted by the computer science, statistical, database, and library and information science communities closely following the advent of the World Wide Web in the following years.

Metadata, which literally means "data about data", refers specifically to descriptive metadata. According to NISO (2004), metadata is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage information." In other words, metadata provides the context in which to interpret data and information. It comprises structured information concerning various entities that can be identified, including but not limited to web pages, books, journal articles, images, songs, products, processes, individuals and their actions, research data, concepts, and services.

Many scholars define the term metadata in different ways.

- "the sum total of what one can say at a given moment about any information object at any level of aggregation" by Gilliland (2016, p.2)
- "Structured, encoded data that describes characteristics of information-bearing entities (including individual objects, collections, or systems) to aid in the identification, discovery, assessment, management, and preservation of the described entities" by Zeng & Qin (2016, p.491)
- "a potentially informative object that describes another potentially informative object" by Pomerantz (2015, p.26)
- "the information we create, store, and share to describe things" by Riley (2017, p.1)

2.2. Types of Metadata

Metadata unlocks the value of data, helping to respond to the "what, where, when, how, and who" of data. There are three main types of metadata: descriptive, administrative, and structural (Riley, 2017).

- i. Descriptive metadata: information that describes the content of the digital object. It enables discovery, identification, and selection of resources. For instance, elements such as title, author, and subjects of a book.
- ii. Structural metadata: information that describes the structure of the digital object. This is generally used in machine processing and describes relationships among various parts of a resource, such as chapters in a book.

- iii. Administrative metadata: information that describes administrative and management aspects of the digital object. It can include elements such as technical, preservation, rights, and usage information.

However, apart from these three main types, there are other types of metadata as follows.

- Technical metadata: information that describes technical aspects of the digital object.
- Preservation metadata: refers to the information related to the preservation management of collections and information resources.
- Provenance metadata: provides helpful information on the origins of a data resource. It includes information on the ownership, any transformation that the data may have undergone and the usage of the data, etc.
- Usage metadata: information collected whenever a user accesses and uses a specific digital information object.
- Legal metadata: provides information about the creator, copyright holder, and public licensing, if provided.

2.3 Metadata Storing, Syntax and Sharing

Metadata are mainly stored in databases often called a metadata registry or metadata repository. In conventional information system architecture, it may be housed within fields within relational database tables. Within this framework, a grouping of metadata is commonly referred to as a record. When importing metadata, there are two ways to do it: in bulk via customized programming or manually via specialized user interfaces. At present, in order to exchange metadata with other systems, software adheres to a metadata model and uses Application Programming Interfaces (APIs). These systems publish specification documents that external software developers can use to create tools capable of retrieving the desired metadata (Riley, 2017).

Metadata syntax is a set of rules created to organize and structure metadata fields or elements. A particular metadata scheme can be expressed in various markup or programming languages, each requiring a different syntax. For example, Dublin Core can be expressed in plain text, HTML, XML, and RDF.

- XML (eXtensible Markup Language)

During the 2000s, XML became a widely used mechanism for encoding, transferring, and occasionally storing metadata within internal systems. In XML, metadata is represented as sets of files, known as XML documents. XML defines elements, which are essentially tags that indicate the meaning of the values contained within them. The utility of XML extends beyond descriptive metadata; a wide array of metadata types can be accommodated within XML documents.

- RDF (Resource Description Framework)

Originally developed as a data model for metadata, RDF has evolved into a widespread method for describing and exchanging graph data. RDF is a standard model for data interchange on the Web. RDF has functionalities that support data

merging even when the underlying schemas are different. It also allows for schema evolution over time without requiring changes to all data consumers. By utilizing URIs for naming relationships between entities, as well as both ends of the link (commonly known as a triple), RDF extends the linking structure of the Web. This enables the integration, exposure, and sharing of structured and semi-structured data across diverse applications (W3C, 2014).

Linked Data is a set of design principles for sharing machine-readable interlinked data on the Web. When combined with Open Data (data that can be freely used and distributed), it is called Linked Open Data (LOD). The concept of Linked Data was introduced in 2006 by Tim Berners-Lee, often recognized as the inventor of the World Wide Web. Linked Data in operation relies heavily on RDF standards. As it builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers.

Fig. 1 shows two separate classes about BOB and Mona Lisa (Subject). Each class bears attributes/values (Object) connected by properties (Predicate). In Linked Data, these statements are often called triples. We can visualize triples as a connected graph. Likewise, any information can be interlinked using this RDF triple technology resulting in a widespread knowledge graph.

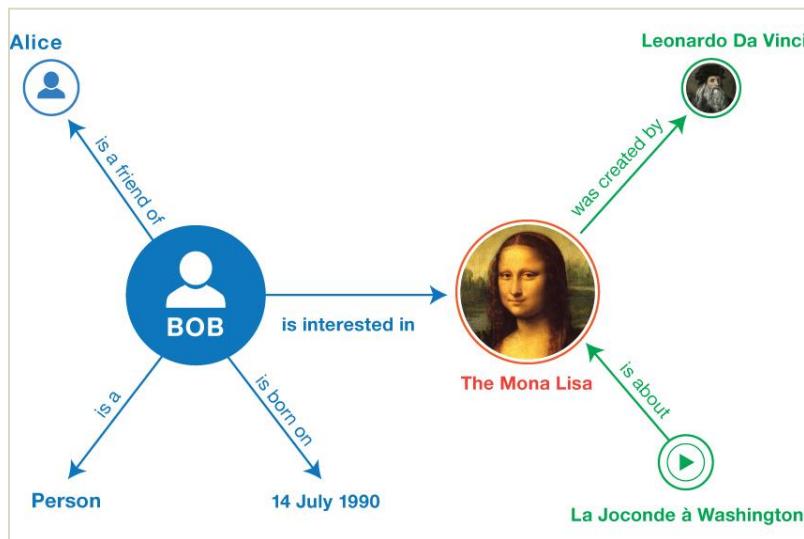


Fig. 1 Informal graph of the sample triples (W3C, 2014)

2.4 Metadata Standards and Tools

i. Dublin Core

Dublin Core comprises fifteen core metadata elements (refer to Fig. 2) utilized for describing digital or physical resources. The Dublin Core Metadata Initiative (DCMI) serves as the primary entity responsible for establishing the Dublin Core standards and

associated specifications. DCMI fosters collaborative innovation in metadata design and promotes best practices across diverse purposes and business models.

Resources described using Dublin Core encompasses digital assets (such as videos, images and web pages) as well as physical items like books or artworks. Dublin Core metadata serves a multitude of purposes, ranging from basic feature descriptions to the integration of metadata vocabularies from various patterns. This facilitates interoperability within the linked data cloud and Semantic Web implementations. Dublin Core applications employ XML and RDF, enhancing accessibility and usability (DCMI, 2024).

Since its straightforwardness, Dublin core elements are extensively utilized by memory institutions and various other organizations to organize their information.

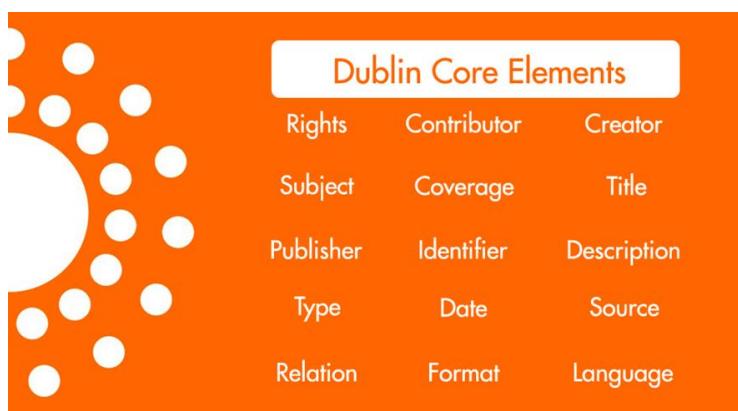


Fig. 2 Dublin Core Elements (Vivian, 2015)

ii. MAchine Readable Cataloging (MARC)

MARC represents a standardized collection of digital formats utilized for machine-readable descriptions of cataloged items in libraries, encompassing books, and digital resources. Automated library catalogs and library management software adhere to the MARC standard to organize their catalog records uniformly across the industry. This facilitates the seamless sharing of bibliographic information between computers. The MARC standard is widely adopted by libraries and other similar institutions globally (Library of Congress, 2023 a).

iii. Metadata Encoding and Transmission Standard (METS)

The METS schema serves as a standard framework for encoding descriptive, administrative, and structural metadata about objects within a digital library. This metadata is articulated using the XML schema language of the World Wide Web Consortium (W3C). Oversight of the standard is carried out by the METS Board in partnership with the Network Development and MARC Standards Office of the Library of Congress. Originating as an initiative of the Digital Library Federation, the METS

schema has evolved into a fundamental tool for organizing and managing digital library resources (Library of Congress, 2023 b).

iv. Metadata Object Description Schema (MODS)

The Metadata Object Description Schema (MODS) is a schema designed for a bibliographic element set that finds utility across various applications, with a particular emphasis on library applications. Maintenance of this standard is carried out by the Network Development and MARC Standards Office of the Library of Congress, ensuring its continued maintenance and relevance within the field of bibliographic metadata management (Library of Congress, 2023 c).

v. VRA Core (Visual Resources Association)

It defines the fields for describing works of visual culture as well as the images which document them. VRA Core is uniquely able to capture descriptive information about both the work and the image and indicate relationships between them (Library of Congress, 2022).

vi. FGDC metadata standards

The Federal Geographic Data Committee (FGDC) is a committee dedicated to fostering the coordinated development, utilization, sharing, and dissemination of geospatial data at a national level in the United States. Among its core responsibilities, FGDC plays a pivotal role in formulating and supporting the adoption, development, and distribution of metadata standards.

FGDC were responsible for the metadata standards program called the Content Standard for Digital Geospatial Metadata (CSDGM) which has been a long-standing metadata standard that is used by many organizations. However, FGDC currently advocates for transitioning metadata standards towards the International Organization for Standardization (ISO) metadata standards (FGDC, 2024).

In this paper, the authors have listed a few metadata standards only. Apart, various standards and models have been developed based on subject domains and requirements. In addition, we can find various data value standards; mainly controlled vocabularies, thesauruses such as the Getty Vocabularies, Library Congress Subject Headings (LCSH), Virtual International Authority File (VIAF), AGROVOC Multilingual Agricultural Thesaurus, Medical Subject Headings (MeSH), and Data Documentation Initiative (DDI) Controlled Vocabularies, etc.

3. Metadata Best Practices in Brazil and Sri Lanka

3.1 Best Practices in Brazil

Metadata is widely used in libraries to describe and catalog books and other materials. They provide information about the content, structure, and context of materials, allowing users to easily find what they are looking for.

There are several initiatives to use metadata in Brazil. One of them is LexML (<https://www.lexml.gov.br/>), which has as its object legislative and legal documents, databases describing the documents (metadata repositories) and document repositories in digital media. Another initiative is the Brazilian Open Data Portal (<https://dados.gov.br/>), which provides datasets on health, transportation, public safety, education, government spending, and the electoral process. Metadata is used to describe and catalog these datasets. The National Library of Brazil uses metadata to describe its collections of photographs in its digital environment (Bettencourt, 2011)

An IBGE initiative is the Geospatial Metadata Profile of Brazil (MGB Profile), which aims to establish a common structure for the description of Geo-information produced in Brazil. It was developed in partnership with the Brazilian defense force and meets the latest international reference standard, ISO 19115-1:201412. The MGB Profile aims to standardize the geospatial metadata available in the National Spatial Data Infrastructure (INDE) (<https://inde.gov.br/>). This allows the cataloguing, integration and harmonization of data produced or maintained and managed in Brazilian government institutions. With this, the producers of geoinformation in the country now have a document with the most current standardization of geospatial metadata, facilitating the search and exploration of geospatial data, and promoting its documentation, integration, and availability. In addition, the MGB Profile is adapted to the national reality, based on the experiences achieved in more than 10 years of INDE's existence (Santos et al., 2015).

In Brazil, Dublin Core is used for cataloging and representation of metadata in libraries and digital repositories, information retrieval, Semantic Web, Linked Data, and interoperability, among other subjects. There is research on the development of the Dublin Core standard and its application in different contexts. An example is a case study conducted at the University of Brasilia that investigated the use of Dublin Core in the representation of information objects in multimedia in the Dulcina de Moraes collection. The study concluded that the application of Dublin Core is feasible to represent and catalog multimedia materials, respecting the limitations of Dublin Core in relation to the specificities of information objects in multimedia (Sousa, 2022).

We found the use of Dublin Core in digital repositories for cataloging and metadata representation. For example, a study conducted by the University of São Paulo investigated the current landscape of institutional repositories of higher education institutions in Brazil and found that the Dublin Core standard is adopted in all Dspace software applications.

3.2 Best Practices in Sri Lanka

MARC (MAchine-Readable Cataloging) is a standard that bridges conventional library catalogues and the machine-readable format of that catalog data. Since libraries have evolved their standard printed catalog to online retrieval systems such as OPAC (Online Public Access Catalog) MARC has become the most popular and widely used standard to retrieve bibliographic details in libraries. Since MARC encodes metadata about bibliographic items it is known as one of the well-known metadata standards as well.

MARC 21 Format for Bibliographic Data contains format specifications for encoding data elements needed to describe, retrieve, and control various forms of bibliographic material such as books, serials, computer files, maps, music, visual materials, and mixed material (Library of Congress, 2007).

Sri Lankan libraries are also embraced with MARC 21, and the majority of the libraries are using the MARC format in their library management systems. Currently, most Sri Lankan libraries are using the Koha Integrated Library Management System built based on MARC 21 format. Additionally, libraries use other MARC-compatible software such as LIBSYS, ALICE for Windows, etc. The number of ‘tags’ in MARC 21 is around 999 providing the users to define more specific metadata related to their resources. Sometimes this advantage itself creates challenging situations while entering metadata into the MARC format. For instance, a certain cataloger may need only a few MARC ‘tags’ to define their resources (Smith-Yoshimura et al., 2010). Also, catalogers might not be aware of which ‘tag’ should be used and which should be omitted. Specially, when forming a Union Catalogue to aggregate several databases identifying a suitable set of ‘tags’ is very crucial.

Therefore, the National Library and Documentation Services Board (NLDSB) of Sri Lanka developed a framework called Descriptive Cataloguing Framework (DCF) to select a set of MARC 21 ‘tags’ that can be utilized by the Sri Lankan libraries. This work was a joint project of the NLDSB and some experts related to the industry in Sri Lanka. The main aim of the project is to facilitate data sharing via a Union Catalogue among Sri Lankan libraries. This aim was based on the following objectives (National Library of Sri Lanka, 2020)

- i. To provide a user guide (framework) for ‘selecting tags’ on MARC for libraries.
- ii. To maintain the uniformity of bibliographic metadata in the county.
- iii. To encourage all automated libraries to utilize the above framework during metadata creation.
- iv. To facilitate collaboration and data sharing

Since the DCF framework is a common format based on 8 areas of description by the ISBD (International Standard Bibliographic Description) this cannot be considered as a broader or complete guideline for libraries. For instance, a library with special intentions can utilize this framework and adapt based on their requirements. Version 01 of the DFC

framework is mainly designated for identifying MARC ‘tags’ related to monographs, books, reports and conference proceedings, etc. The NLDSB is continuing its efforts to create frameworks for serials, articles, book chapters and manuscripts subsequently. In addition, NLDSB is currently planning to develop some suitable metadata framework for the Palm Leaf Manuscripts of Sri Lanka. Hopefully it will be another good practice which enables us to share and preserve resources related to cultural heritage in Sri Lanka.

The second recent best practice related to metadata initiatives in Sri Lanka is rather different from the above scenario. This initiation is called National Spatial Data Infrastructure (NSDI) Sri Lanka initiated by the Information and Communication Technology Agency of Sri Lanka (ICTA) with collaboration of many stakeholders who utilize geospatial data in Sri Lanka. NSDI has been identified as one of the key initiatives under the National Digital Policy - “Digitalization of the Economy” by the Government of Sri Lanka (NSDI, 2022). The objectives of an NSDI are to provide a platform to create, analyze and discover spatial information online for effective evidence-based decision-making.

This project was initiated in 2016, focusing on creating standard infrastructure and solutions; to avoid data duplication, improve data quality, standardize spatial data, improve transparency in data sharing across institutions and provide a technology platform for developing spatial data decision support tools while collaborating with various government and non-government stakeholders. Currently, the project has accomplished a National Map Portal, a Metadata Catalogue, some use cases of NSDI and a web portal (<https://nsdi.gov.lk>), etc. In addition they are currently developing certain policies such as National Data Sharing Policy, etc.

4. Discussion and Conclusions

Metadata is part of the structure of a main piece of data, being responsible for making it a set of useful information. They are used to aid in cataloguing and retrieving data, forming the foundation of the semantic web; an extension of the web that allows humans and machines to interact through code.

In this paper, the authors have introduced basic metadata conceptions including the definitions of metadata. Furthermore, it gave some idea on core metadata standards.

In section 3, the case studies from two different countries were presented. Brazil and Sri Lanka both have some initiatives related to library and Geospatial data domains. The scale of these projects may be different but we can see some similarities in both scenarios.

Libraries use various metadata standards to organize and retrieve resources. This applies to both countries. Compared to Brazil, Sri Lankan Library, Archives and Museum’s (LAM) awareness of metadata is confined to MARC standards. Similar to Brazil, many institutions in Sri Lanka are maintaining Dspace repositories which use a Qualified Dublin Core-based metadata schema. Some professionals in the information science sector are aware of Dublin Core standards, but the use cases are very rare. Based on the above best practices, awareness of Dublin Core metadata standards in Brazil is

wider compared to Sri Lanka. However, this scenario should be further investigated thoroughly.

In addition, the Brazilian Open Data Portal can be identified as another decent example of best practices in order to access public data in Brazil. Open and Linked data are closely connected with the metadata domain and this specific example has used these technologies and different metadata standards while developing their portal.

Finally, awareness of basic conceptions and standards related to metadata is a key factor when it comes to conducting more research in the domain. Currently, LIS (Library & Information Science) students are taught about metadata standards during their postgraduate programs. Apart, regular webinars, workshops are conducted by universities and institutions such as the National Library of Sri Lanka and Brazil to teach and enlighten people of the use of metadata in different domains.

References

Bettencourt, A. M. M. (2011). A representação da informação na Biblioteca Nacional do Brasil: do documento tradicional ao digital.

Baca, M. (Ed.). (2016). Introduction to metadata. 3rd Ed., Los Angeles: Getty Research Institute.
<https://www.getty.edu/publications/intrometadata/>

DCMI. (2024). About DCMI. <https://www.dublincore.org/about/>

FGDC. (2024). Geospatial Metadata Standards and Guidelines.
<https://www.fgdc.gov/metadata/geospatial-metadata-standards>

Gilliland, A.J. (2016). Setting the stage. In M. Baca (Ed.), Introduction to metadata (3rd ed., pp. 1–20). Los Angeles: Getty Research Institute.

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly, 40:3-4, 17-36. DOI: 10.1300/J104v40n03_02

Iescheck, A. L. & Dorneles, M. A. (nd.). Brazilian National Spatial Data Infrastructure (INDE): Applicability for Large Scale Data.

Lee Eden, B. (2004). Metadata and Librarianship: Will MARC Survive?. Library Hi Tech, 22(1), 6-7.

Library of Congress. (2023 a). MARC Standards. Library of Congress- Network Development and MARC Standards Office. <https://www.loc.gov/marc/>

Library of Congress. (2023 b). Metadata Encoding and Transmission Standard (METS).
<https://www.loc.gov/standards/mets/>

Library of Congress. (2023 c). Metadata Object Description Schema (MODS).
<https://www.loc.gov/standards/mods/>

Library of Congress. (2022). VRA Core: A Data Standard for the Description of Works of Visual Culture. <https://www.loc.gov/standards/vrare/>

Library of Congress. (2014). MARC Standards: Marc 21 Formats.
<https://www.loc.gov/marc/marcdocz.html>

Library of Congress. (2007). The MARC 21 Formats: Background and Principles (Revised November 1996). Library of Congress. <https://www.loc.gov/marc/96principles.html>

National Library of Sri Lanka. (2020). MARC 21 Descriptive Cataloguing Framework Recommended by the National Library of Sri Lanka. NLSL (Version 1.0).
http://www.natlib.lk/pdf/MARC21_DCFSL_V1.pdf

NISO. (2004). Understanding Metadata. National Information Standards Organization (NISO) Press. <https://web.archive.org/web/20141107022958/http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

NSDI. (2022). National Spatial Data Infrastructure (NSDI). Information and Communication Technology Agency of Sri Lanka (ICTA). <https://nsdi.gov.lk/>

Pomerantz, J. (2015). Metadata. Cambridge, MA: MIT Press.

Riley, J. (2017). Understanding Metadata. Washington DC, United States: National Information Standards Organization, 23, 7-10.
<https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf>

Santos, C. J. B. et al. (2015). The National Spatial Data Infrastructure of Brazil (INDE) Making Visible Some Invisible: The Case of the Brazilian Geographical Indications. Brazilian Journal of Cartography. N° 67/5 Special Issue 27th ICC: 1025-1033 Brazilian Society of Cartography.

Smith-Yoshimura, K. et al. (2010). Implications of MARC tag usage on Library Metadata Practices. Report produced by OCLC Research in support of the RLG Partnership.
<https://www.oclc.org/content/dam/research/publications/library/2010/2010-06.pdf>

Sousa, J. P. L. D. (2022). Uso de Dublin Core na representação de objetos de informação em multimeios: um estudo de caso no acervo Dulcina de Moraes [Monografia]. Universidade de Brasília.

Vivian. (2015). Dublin Core vs Schema.org: A Head-To-Head Metadata Comparison. SeoProfessor. <https://seopressor.com/blog/dublin-core-vs-schemaorg-metadata-comparison/>

World Wide Web Consortium (W3C). (2014). RDF 1.1 Primer. W3C Working Group.
<https://www.w3.org/TR/rdf11-primer/>

Zeng, M. L., & Qin, J. (2016). Metadata (2nd ed.). Chicago: ALA.