METHODOLOGICAL PRINCIPLES TO CREATE A METADATA EXTENSION TO THE DARWIN CORE STANDARD FOR AGROBIODIVERSITY DATA

Princípios metodológicos para criar uma extensão de metadados para o padrão Darwin Core para dados de agrobiodiversidade

Filipi Miranda Soares (1), Benildes Coura Moreira dos Santos Maculan (2), Debora Pignatari Drucker (3), Antonio Mauro Saraiva (4)

(1) Escola Politécnica, Universidade de São Paulo, Brazil, filipisoares@usp.br. (2) Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Brazil, benildes@gmail.com. (3) Embrapa, Brazil, debora.drucker@embrapa.br. (4) Escola Politécnica, Universidade de São Paulo, Brazil, saraiva@usp.br

Abstract

This paper aims to propose principles for creating a metadata extension to the Darwin Core standard that addresses the agrobiodiversity data, with a scope on ecological interactions. These principles have been compiled from the scientific literature, giving special attention to recommendations of the DCMI Abstract Model, which outlines the principles for creating metadata. The DCMI Abstract Model governs the creation of the Dublin Core metadata standard upon which Darwin Core is based. The requirements of ISO/IEC 11179-4/2004 standard for the definition of metadata were also taken into consideration. A prototype of a metadata record for the field of ecological interactions, which is the scope of this research within agrobiodiversity, was created to demonstrate the format that metadata will have when the extension is finished. This research an effort to propose more effective tools for agrobiodiversity data management, but it is necessary to mature and deepen the discussions around the conceptual aspects of the ecological interactions in agrobiodiversity and the relationship of the new metadata extension with the vocabulary of the Darwin Core, as well a robust methodology to create DwC extensions is still pending of being developed.

Keywords: Metadata; Metadata extension; Darwin Core; Agrobiodiversity; Information representation

Resumo

Esta pesquisa propõe princípios para se criar uma extensão de metadados para o padrão Darwin Core para representar dados de agrobiodiversidade, com escopo temático nas interações ecológicas. Estes princípios foram compilados a partir da literatura científica, com ênfase nas recomendações do DCMI Abstract Model, que apresenta princípios para criação de metadados. O DCMI Abstract Model rege a criação do padrão Dublin Core, no qual o Darwin Core se baseia. Os requisitos da norma ISO/IEC 11179-4/2004

Soares, Filipi Miranda, Maculan, Benildes Coura Moreira dos Santos, Drucker, Debora Pignatari, and Saraiva, Antonio Mauro. Methodological principles to create a Metadata extension to the Darwin Core standard for agrobiodiversity data. *Brazilian Journal of Information Science:* Research trends, vol. 14, no.4, set.-dez. 2020, e020015 https://doi.org/10.36311/1940-1640.2020.v14n4.10865

para definição de metadados também foram seguidos. Um protótipo de um registro de metadados para o campo das interações ecológicas, que é o escopo da pesquisa dentro da agrobiodiversidade, foi criado para demonstrar o formato que os metadados terão quando a extensão for finalizada. Esta pesquisa representa um esforço para propor ferramentas mais eficazes para a gestão de dados de agrobiodiversidade, mas é necessário amadurecer e aprofundar as discussões em torno dos aspectos conceituais das interações ecológicas na agrobiodiversidade e da relação da nova extensão de metadados com o conjunto de termo do Darwin Core, bem como uma metodologia robusta para criar extensões para o DwC ainda está pendente de ser desenvolvida

Palavras-chave: Metadados; Extensão de metadados; Darwin Core; Agrobiodiversidade; Representação da informação

1 Introduction

Metadata creation and curation are community-driven tasks. Many metadata standards have been developed by scientific communities for distinct knowledge fields. Metadata for specific subjects are named disciplinary metadata. The Digital Curation Center (DDC) presents dozens of disciplinary metadata standards currently in use for all disciplines of knowledge on the website: https://www.dcc.ac.uk/guidance/standards/metadata.

Some metadata standards have been developed for the biodiversity science over the history, such as Access to Biological Collection Data (ABCD), Darwin Core (DwC), and Ecological Metadata Language (EML). Between them, DwC is the most used metadata standard to share data about biodiversity in the Global Biodiversity Information Facility (GBIF) portal[1], one of the largest biodiversity data repositories in the world (Body et al. 2020). Its worldwide use makes us believe that DwC (Wieczorek et al. 2012) may be used to describe agrobiodiversity data. However, pragmatic analysis of DwC and DwC Metadata Extensions demonstrated that important concepts and relations of Agricultural Biodiversity are not represented in DwC elements (Soares et al. 2019; Soares 2019).

The Convention on Biological Diversity (CBD 2000) defines Agricultural Biodiversity as the set of elements of biodiversity that are relevant somehow to agriculture and food production. In other words, "the variability among living organisms from all sources including terrestrial,

[1] Available from: https://www.gbif.org/

marine and other aquatic ecosystems and the ecological complexes of which they are part: this includes diversity within species, between species and of ecosystems" (CBD 2000 p. 85).

The field and research work in Agricultural Biodiversity produces data. The Brazilian Agricultural Research Corporation (Embrapa) is a leading institution in Agriculture research in Brazil. Between the topics investigated by Embrapa, there is Agricultural Biodiversity, such as the occurrence of biological interactions in crops and fields as mechanisms of plague and disease control, soil nutrient cycling, plant nutrition, agroforestry, etc. A big part of the results of the research conducted by Embrapa are published in textual form as journal papers, Ph.D. dissertation or Master thesis, works in event proceedings, books, and book chapters. Just in 2018, Embrapa has published 7.687 documents of the mentioned types (EMBRAPA 2019). Within this context of multiple publications, it is hard to find all the information about biological interactions mapped by Embrapa collaborators, as there is no repository made up with proper metadata to cover specifically this subject.

Given this problem, a research project[2] is going on to develop a metadata extension able to represent data about agricultural biodiversity produced by Embrapa. Nevertheless, before creating a metadata extension, it is necessary to set rules to standardize this process. Thus, this paper aims to present some methodological principles required to create a new metadata extension to the DwC, within the scope of agrobiodiversity data.

2 Information representation and metadata

A representation is a piece of information that describes a digital object in a way it can be retrieved on the web or on a database (Chu 2005). “Information representation includes the extraction of some elements (e.g., keywords or phrases) from a document or the assignment of terms (e.g., descriptors or subject headings) to a document so that its essence can be characterized and presented” (Chu 2005 p. 14).

[2] The project begins at the Federal University of Minas Gerais (UFMG) (Soares 2019) in collaboration with Embrapa as a master’s degree research and is carried on at Polytechnical School of São Paulo University (USP) as a Ph.D. research (Soares et al. 2020).

Soares, Filipi Miranda, Maculan, Benildes Coura Moreira dos Santos, Drucker, Debora Pignatari, and Saraiva, Antonio Mauro. Methodological principles to create a Metadata extension to the Darwin Core standard for agrobiodiversity data. *Brazilian Journal of Information Science:* Research trends, vol. 14, no.4, set.-dez. 2020, e020015 https://doi.org/10.36311/1940-1640.2020.v14n4.10865

Metadata has emerged of the need to organize the growing amount of digital information on the web to improve information retrieval (Alves 2005; 2016).

Information representation is a field of study in Library Science, but also of Informatics and Computing Linguistic (Lourenço 2016). The term metadata has emerged back in the 1960s and was applied to the bibliographic description in libraries, but became popular just in 1995 with the emergence of Dublin Core metadata standard, created for describing digital objects on the web (Alves 2016; Lourenço 2016).

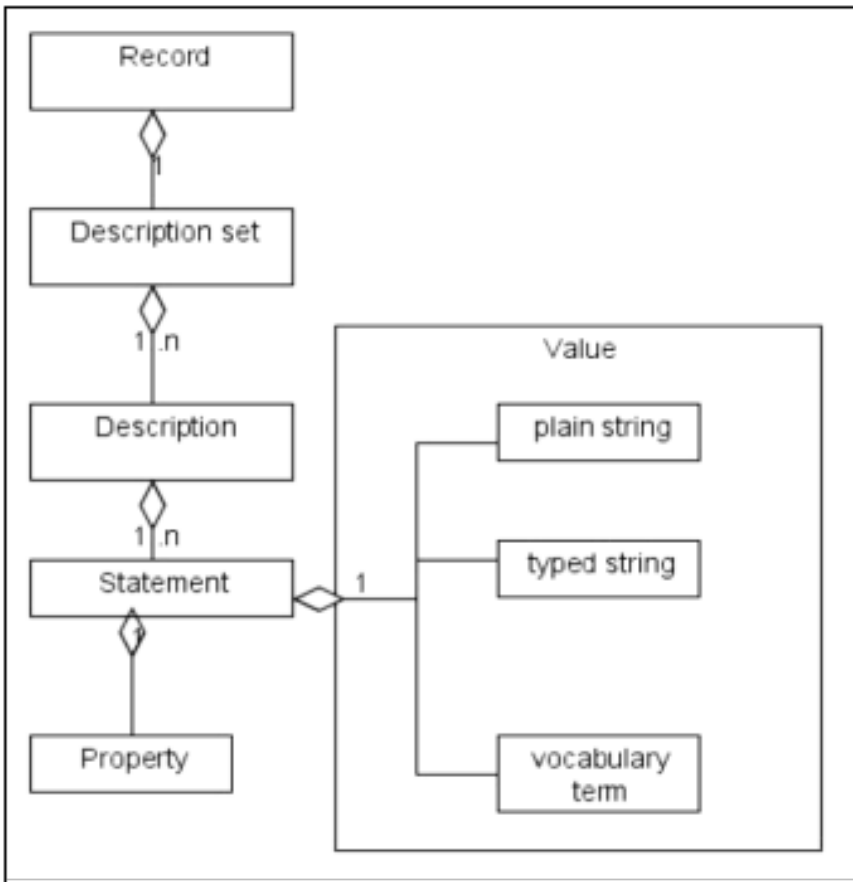
The main aim of information representation is to enable information retrieval (Lourenço 2016). To reach this propose, many models and definitions have been developed to support metadata creation and use.

2.1 Metadata

Metadata may be understood as labels created to describe data content (National Information Standards Organization 2007; Pomerantz 2015; Riley 2017; Zeng 2015; Zeng and Qin 2008). It Is Often Defined as "data about data" in literature, e.g. in the ISO/IEC 11179-4 (2004) standard for metadata registries, but this trivial definition may not be enough (Pomerantz 2015). However, there are definitions in the literature that better express the whole function of metadata. Zeng (2015) indicates the variations of the definition for the metadata concept through different communities of practice but shows a definition that better fits within the research approach of this paper: metadata are "information about specific things" (Zeng 2015). This definition by Zeng (2015) seems to be more adequate than the one of ISO/IEC 11179-4 (2004). However, the definition that better expresses the meaning of the concept is given by the National Information Standards Organization (National Information Standards Organization 2004 p. 1): "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information".

E. g., consider the number 5531986424933. It is hard to understand what this number means without any contextual information. However, if the number is displayed in the following way, it is much easier to understand: "<phone>5531986424933</phone>. The label phone describes that set of data as a telephone number.

Figure 1 – DCMI Abstract Model simple



Source: Coyle (2008)

Dublin Core (DC) is the first metadata standard created to describe information resources of any subject on the Web. The Dublin Core Metadata Initiative (DCMI), the group responsible for the development and maintenance of the DC, set up some principles that may help to create metadata for any subject field, known as the DCMI Abstract Model (Powel et al. 2007).

The DCMI Abstract Model is an overly complex semantic structure, but it is possible to make it simple when creating new metadata. The simple version of the DCMI Abstract Model was first presented by Coyle (2008) as Figure 1.

Figure 1 shows that a metadata record is a set of descriptions, and this set may have one or many descriptions; the description is one or many statements made about the subject represented in the metadata record; the statement has both a property and a value, part of a triple (subject-predicate-object). A predicate is always a term that describes the data (value). Taking up

the example from before, <phone> is a predicate, 5531986424933 is a value and it is a phone number that belongs to a person — the subject in the Triple. This model is very similar to the Resource Description Framework (RDF), which is "a globally-accepted framework for data and knowledge representation that is intended to be read and interpreted by machines" (Wilkinson et al. 2016 p. 2).

> RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. (RDF Working Group 2014).

All this effort has a much bigger aim: making scientific data FAIR: Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016).

Table 1 – Darwin Core element eventTime

eventTime	**Property**
Identifier	http://rs.tdwg.org/dwc/terms/eventTime
Definition	The time or interval during which an Event occurred.
Comments	Recommended best practice is to use a date that conforms to ISO 8601-1:2019.
ExampleS	14:07-0600 (2:07pm in the time zone six hours earlier than UTC). 08:40:21Z (8:40:21am UTC). 13:00:00Z/15:30:00Z (the interval between 1pm UTC and 3:30pm UTC).

Source: Darwin Core Task Group (2009d)

To make scientific data FAIR, the metadata standards are made not just of triples or elements. It also requires a set of clear and well-defined rules. "Any metadata item that is to be retrieved directly (as opposed to indirectly through a related item), shall be an identified item, so the item can be referenced. Each identified item shall have at least one identifier, and that identifier must be unique within a specified namespace" (ISO/IEC 2015 p. 18). I.e., ISO/IEC

11179-6 (2015) shows that each element in the metadata schema must have a unique identifier. Beyond that, to create a metadata schema one must define the semantic and syntax rules for it. The semantic gives the meaning of each element of the metadata schema, setting its function on the registry, while the syntax determines how to format the metadata in an interoperable way. Observe the DwC element <eventTime> in Table 1.

The attribute *identifier* gives the pathway to a computer program to understand the element (makes it interoperable); *definition* in Table 1 sets the meaning of the metadata element, i.e., its semantics. Meanwhile, *comments* set part of the syntax of the metadata element by indicating an encoding scheme to properly format data. The *examples* show how the data look like if properly formatted. Each metadata standard sets its own rules, e.g. a set of elements that must be always present in the metadata record or how to organize the metadata classes of the metadata record. DwC set of rules for formatting and using metadata can be found in the documentation published by the Darwin Core Task Group (2009a; 2009b; 2009c; 2009d) and Darwin Core and RDF/OWL Task Groups (2015).

3 Agrobiodiversity data

A report published by the GBIF task group on data fitness for use in agrobiodiversity (Arnaud et al. 2016) shows the need for developing strategies and tools to manage agrobiodiversity data. Arnaud et al. (2016) point out that the focus on agrobiodiversity data might be on taxon, vernacular names, occurrences, geospatial distribution, genotype, phenotype, environmental factors, agronomic traits, functional traits, species interactions, socioeconomic factors, and local knowledge. There are metadata in DwC standard that describe taxon, vernacular names, occurrences, and geospatial distribution; in DwC metadata extensions, it is possible to find metadata for genetic data; the other concepts are uncovered by DwC metadata. Thus, this research focuses on species interactions, a subject field unexplored in the scope of DwC.

Species interact all the time in crops and farms, so knowing those interactions is particularly important for food production. E. g., the pollination, a kind of mutualistic interaction between an animal (e.g. a bee, bird, or a bat) and a plant, is crucial to produce fruits and seeds.

According to the Food and Agriculture Organization of the United Nations (FAO 2018 p. 3) "three out of four crops across the globe" depend on pollinators to reach yields. Beyond pollination, many other interactions are noticed in agriculture. These interactions are presented as a conceptual model in Figure 2 in section 5.1.

4 Methods

The methodological principles we believe are necessary to create a DwC metadata extension were assembled in three phases:

a) selection and analysis of terminological and data inputs;

b) terminological definition and metadata modeling;

c) the community of practice evaluation.

4.1 Selection and analysis of terminological and data inputs

This stage of the methodology aimed the immersion in the thematic field of agrobiodiversity. It was organized into five sub-steps:

a) definition of the scope of agrobiodiversity data representation, using as guide the Final Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity (Arnaud et al. 2016), to set the sample to be worked (ecological interactions);

b) literature analysis to set definitions of ecological interactions' concepts;

c) analysis of DwC terms;

d) analysis of DwC metadata extensions;

e) compilation of principles from ISO/IEC 11179-4 (2004), RDF Scheme, and DwC documentation.

4.2 Terminological definition and metadata modeling

Metadata are represented by predicates, which are terms that have meaning and a representative function. The function of metadata can only be understood through clear and well-established definitions, plus an International Resource Identifier (IRI) or a Uniform Resource

Identifier (URI), which make the metadata element unique and avoid misunderstanding for both humans and machines.

This second step consisted of applying principles that allow us to define, clearly and objectively, the function of metadata elements. Those are ISO/IEC 11179-4 (2004) recommendations for metadata construction and the syntactic and semantic scheme of DwC, based on RDF schema.

This task of defining metadata terms has been named 'functional terminological definition', since this activity consists in establishing the function of metadata, pointing out its rules of application.

5 Results and analysis

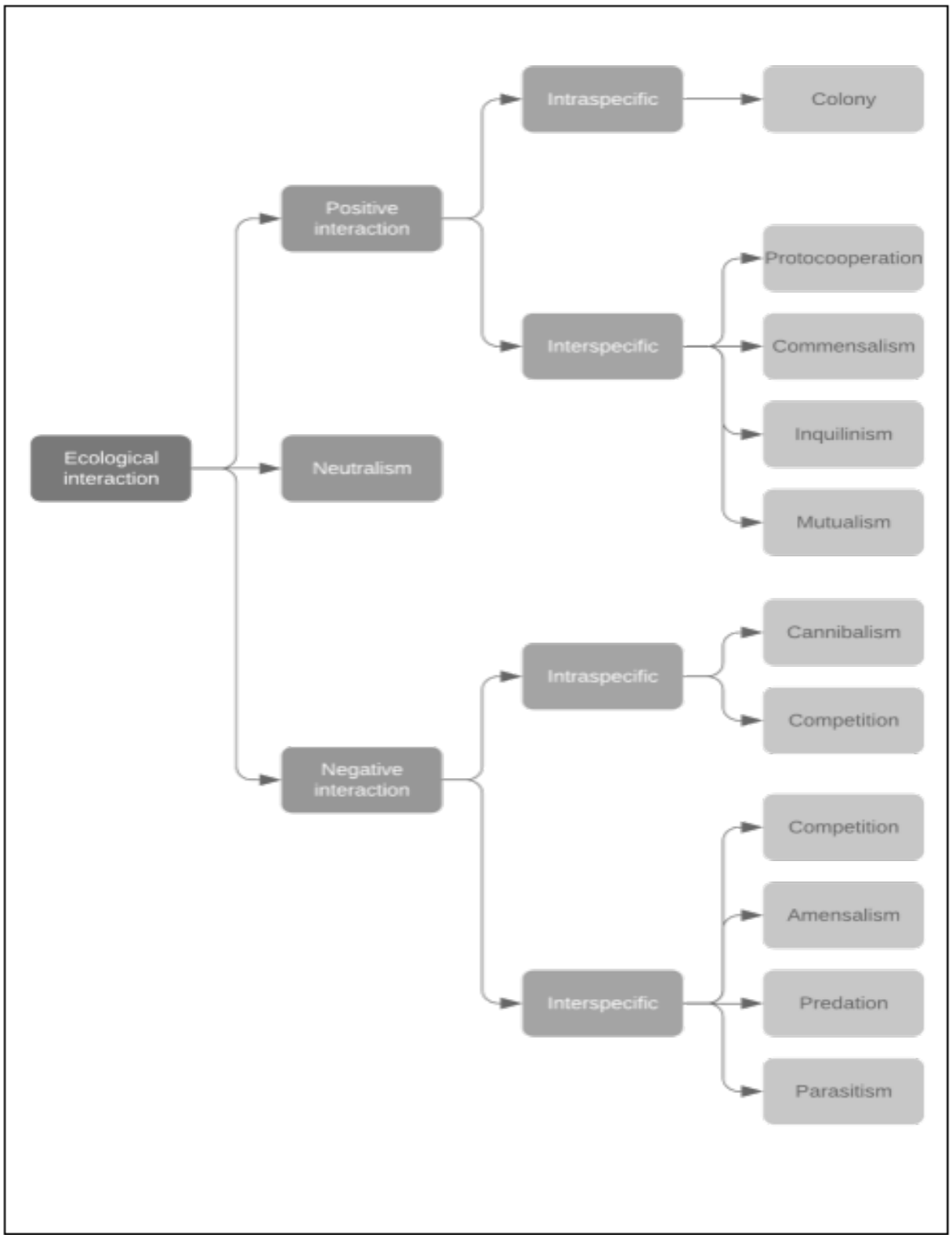
In this section, we present the principles we followed to start the task of creating the metadata extension.

5.1 Analysis of methodological and terminological inputs

The preview analysis of species interaction concepts in the literature (Cain et al. 2017; Cassini 2005; Odum and Barret 2005; Pinto-Coelho 2007; Stein 2018) resulted in the model of Figure 2, which covers the broader concepts. A deeper investigation of the agrobiodiversity literature is being executed in the second stage of this research project to find out the most frequent species interactions in agricultural ecosystems, as crops and fields. The model of Figure 2 shows a concept schema for the interaction grid, a model that represents species interactions.

An ecological interaction is an action that involves two organisms of the same species (intraspecific interaction) or two different species (interspecific interaction). Thus, to represent this relationship, one must define the role of each organism or species in the interaction. For example, in predation, one organism is a predator, which means it eats another living being. The other organism which serves as food for the predator is denominated prey. Each organism, if represented by metadata, is a resource. The terms “predator” and “prey” are the whole of each organism in the interaction and may be useful to understand the relationship between the resources.

Figure 2 – Species interactions



Source: elaborated by the authors (2020)

Just analyzing the literature is not enough to define metadata to describe the ecological interactions in agrobiodiversity data. For an information scientist or a computer scientist, this literature analysis enables them to understand the basic concepts of ecological interactions agrobiodiversity, but not more than that. It is mandatory to get involved with agrobiodiversity specialists who will say what data matters to be represented.

5.2 Terminological definition and metadata modeling

The definition of the metadata terms must establish the function of the metadata. This definition is different of a mere definition of a glossary or a dictionary term: it must define the semantics and the syntax of the metadata element, which determine the role that the metadata element will play in the data description. For example, observe the following two definitions of the word 'date', the first one as a metadata element of Dublin Core and the second one as a dictionary definition:

> a) dc:date: "A point or period of time associated with an event in the lifecycle of the resource. Comment: Date may be used to express temporal information at any level of granularity. Recommended practice is to express the date, date/time, or period of time according to ISO 8601-1 [ISO 8601-1] or a published profile of the ISO standard, such as the W3C Note on Date and Time Formats [W3CDTF] or the Extended Date/Time Format Specification [EDTF]. If the full date is unknown, month and year (YYYY-MM) or just year (YYYY) may be used. Date ranges may be specified using ISO 8601 period of time specification in which start and end dates are separated by a '/' (slash) character. Either the start or end date may be missing." (DCMI Usage Board 2020);
>
> b) date (Oxford Learner's Dictionaries): "*noun*; particular day/year; 1) a particular day of the month, sometimes in a particular year, given in numbers and words; 2) a particular day or year when a particular event happened or will happen; 3) a time in the past or future that is not a particular day; 4) an arrangement to meet somebody at a particular time; 5) a meeting that you have arranged with a boyfriend or girlfriend or with somebody who might become a boyfriend or girlfriend; 6) a boyfriend or girlfriend with whom you have arranged a date; 7) a sweet sticky brown fruit that grows on a tree called a date palm, common in North Africa and West Asia". (Oxford University Press 2020).

As we see, the first definition given by Dublin Core is objective: it says that date is a period of time and recommends using ISO 8601-1, W3CDTF or EDTF to express the date. In other words, it gives the guidelines to format the data of the type 'date' in any metadata record, so it can be more easily retrieved by computer applications and shared with other information systems. In the definition given by the Oxford Learner's Dictionaries it is possible to note many

different definitions for the word ‘date’ that are not related to a period of time: an arrangement, a meeting with a boyfriend or girlfriend (or with somebody who might become a boyfriend or girlfriend), a sweet sticky brown fruit. A metadata term must avoid this multiplicity of interpretations — what makes it different of a dictionary term because the information systems and computer applications must “talk the same language” when they are sharing information. To do so, ISO/IEC 11179-4 (2004) requirements give the guidelines to standardize the definition of metadata elements (see Table 2).

Table 2 – ISO/IEC 11179-4 requirements for metadata creation

REQUIREMENTS		
be stated in the singular	EXPLANATION	The concept expressed by the data definition shall be expressed in the singular. (An exception is made if the concept itself is plural.)
	EXAMPLES	Article Number: a) **good definition**: a reference number that identifies an article; b) **poor definition**: reference number identifying articles.
	REASON	The poor definition uses the plural word “articles,” which is ambiguous since it could imply that an “article number” refers to more than one article.
state what the concept is, not only what it is not	EXPLANATION	When constructing definitions, the concept cannot be defined exclusively by stating what the concept is not.
	EXAMPLES	Freight Cost Amount: a) **good definition**: cost amount incurred by a shipper in moving goods from one place to another; b) **poor definition**: costs which are not related to packing, documentation, loading, unloading, and insurance.
	REASON	The poor definition does not specify what is included in the meaning of the data.
be stated as a descriptive phrase or sentence(s) (in most languages)	EXPLANATION	A phrase is necessary (in most languages) to form a precise definition that includes the essential characteristics of the concept. Simply stating one or more synonym(s) is insufficient. Simply restating the words of the name in a different order is insufficient. If more than a descriptive phrase is needed, use complete, grammatically correct sentences.
	EXAMPLES	Agent Name: a) good definition: name of party authorized to act on behalf of another party; b) poor definition: representative.
	REASON	“Representative” is a near-synonym of the data element name, which is not adequate for a definition.

Table 2 – ISO/IEC 11179-4 requirements for metadata creation (continued)

be stated as a descriptive phrase or sentence(s) (in most languages)	EXPLANATION	A phrase is necessary (in most languages) to form a precise definition that includes the essential characteristics of the concept. Simply stating one or more synonym(s) is insufficient. Simply restating the words of the name in a different order is insufficient. If more than a descriptive phrase is needed, use complete, grammatically correct sentences.
	EXAMPLES	Agent Name: a) **good definition**: name of party authorized to act on behalf of another party; b) **poor definition**: representative.
	REASON	"Representative" is a near-synonym of the data element name, which is not adequate for a definition.
contain only commonly understood abbreviations	EXPLANATION	Understanding the meaning of an abbreviation, including acronyms and inicialisms, is usually confined to a certain environment. In other environments, the same abbreviation can cause misinterpretation or confusion. Therefore, to avoid ambiguity, full words, not abbreviations, shall be used in the definition. Exceptions to this requirement may be made if an abbreviation is commonly understood such as "i.e." and "e.g." or if an abbreviation is more readily understood than the full form of a complex term and has been adopted as a term in its own right such as "radar" standing for "radio detecting and ranging." All acronyms must be expanded on the first occurrence.
	EXAMPLES	Tide Height: a) **good definition**: the vertical distance from mean sea level (MSL) to a specific tide level; b) **poor definition**: the vertical distance from MSL to a specific tide level.
	REASON	The poor definition is unclear because the acronym, MSL, is not commonly understood and some users may need to refer to other sources to determine what it represents. Without the full word, finding the term in a glossary may be difficult or impossible.

Table 2 – ISO/IEC 11179-4 requirements for metadata creation (continued)

be expressed without embedding definitions of other data or underlying concepts	EXPLANATION	As shown in the following example, the definition of a second data element or related concept should not appear in the definition proper of the primary data element. Definitions of terms should be provided in an associated glossary. If the second definition is necessary, it may be attached by a note at the end of the primary definition's main text or as a separate entry in the dictionary. Related definitions can be accessed through relational attributes (e.g., cross-reference).
	EXAMPLES	Sample Type Code: a) **good definition**: a code identifying the kind of sample; b) **poor definition**: a code identifying the kind of sample collected. A sample is a small specimen taken for testing. It can be either an actual sample for testing or a quality control surrogate sample. A quality control sample is a surrogate sample taken to verify the results of actual samples.
	REASON	The poor definition contains two extraneous definitions embedded in it. They are definitions of "sample" and of "quality control sample."

Source: ISO/IEC 11179-4 (2004 p. 4-6)

These ISO/IEC principles are useful if one does not know from where to start to create metadata. However, some of its principles are not very up to date to the practice of metadata creation for specific scientific data fields. For example, one ISO/IEC rules declare that the definition of a metadata term should not embed definitions of other metadata elements. It helps to avoid unnecessary repetitions, but sometimes a cross-reference definition is needed to complement the meaning of the metadata element, especially when these elements are arranged into classes. For example, in DwC (Darwin Core Task Group 2009d) the property <kingdom> has the definition "The full scientific name of the kingdom in which the taxon is classified". The scientific name is another metadata property of DwC, but its name is embedded in <kingdom> property definition to show what kind of data this element can represent.

5.3 Shaping the metadata extension

To illustrate the application of phases two and three of the methodology, we present the following situation: we wish to describe the ecological interaction between two species used in biological control. "Biological control can be defined as using living natural enemies to control pests (Kenis et al. 2019 p. 1). To control other living organisms' population, the organism used

as biological control must be a predator or a parasite of the species that one wants to control or to exterminate.

In a research conducted at Embrapa, Pacheco and Corrêa-Ferreira (2000) presented the use of a species of the wasp *Telenomus podisi* as a parasite to control the population of three species of stink bug (*Euschistus heros, Piezodorus guildinii* & *Nezara viridula*) that attack soybean crops.

Given this situation, this interaction classified as Parasitism can be represented in a triple as the diagram in Figura 3.

Figure 3 – Parasitism interaction in the triple model

Source: elaborated by the authors (2020)

The predicate *parasite of*, which describes the relationship between the wasp *Telenomus podisi* (subject) and the *Euschistus heros* stink bug (object) in Figure 3, can be represented as a DwC metadata property as follows in Table 3.

Table 3 – Element parasiteOf

parasiteOf	**Property**
Identifier	https://git.io/JT2fM
Definition	A living being that lives at the expense of another living being, harming it.
Comments	The full scientific name, with authorship and date information if known, of the host species.
Example	*Euschistus heros* (Fabricius, 1794)

Source: Soares (2019)

The *identifier* in Table 3 is a reference to the GitHub repository where the metadata element is published. The *definition* is based on Sorci and Garnier (2008). The label parasite was imported from AGROVOC[3] which presents the term parasites in plural form. However, the term was adopted in the singular form to meet the first requirement of ISO/IEC 11179-4 (2004) in Table 2.

The definition for the property in Table 3 applies the basic requirements of ISO/IEC 11179-4 (2004) for metadata construction:

a) the term used as metadata element is stated in the singular ('parasite', not 'parasites');

b) the definition states what the concept is, in fact, not just what it is not;

c) the definition sentence is descriptive, that is, it clarifies the meaning of the term used as a metadata element;

d) it does not contain abbreviations difficult to understand;

e) it does not embed definitions of other metadata or underlying concepts.

Table 4 – Element hostOf

hostTo	**Property**
Identifier	https://git.io/JT2U3
Definition	A living being that provides shelter for another living being.
Comments	The full scientific name, with authorship and date information if known of the parasitic species.
Examples	*Telenomus podisi* (Ashmead 1893)

Source: Soares (2019)

Considering that the object of the example in Figure 3, that is, the stink bug species *Euschistus heros* (Fabricius 1794) can also be the subject (a resource) of a metadata record if the

[3] Available from: http://aims.fao.org/aos/agrovoc/c_5574. Access on: 24 Nov. 2019.

triple is inverted, another predicate can be created, with the stink bug as the subject, as shown in Table 4.

The *identifier* in Table 4 is a reference to the GitHub repository where the metadata element is published. The *definition* is based on Sorci and Garnier (2008). The host label was imported from AGROVOC[4] which presents the term hosts in the plural. Nevertheless, the term was adopted in the singular to meet the first requirement of ISO/IEC 11179-4 (2004) in Table 2.

The element in Table 4 follows the same rules of specification as in the element in Table 3. However, the nature of this element is more complex: an organism can be a host in four kinds of ecological interactions shown in Figure 2: parasitism, commensalism, inquilinism, and mutualism. In each of these interactions, the function of the host varies: in parasitism, the host is harmed by the parasite, so it is called a negative interaction, but it is not harmed in commensalism; mutualism or inquilinism, called positive interactions. It is important to make explicit the role of the organism in the interaction within the metadata record, as it determines if a given organism can be used as a resource of agrobiodiversity in crops and farms, or not.

The same species may be involved in one or more interactions, so it may assume the role of the host more than once in different interactions. This implies repeating the hostOf element in the metadata record, which Simple DwC usage rules recommend not to do (Darwin Core Task Group 2009b). A possible solution to this problem is to organize the metadata into classes, using RDF model since it has no limitation on repeating properties (Darwin Core and RDF/OWL Task Groups 2015). The hostOf element subordinate to the Parasitism class in a metadata record ceases to be ambiguous: it becomes apparent that it is a host that houses a parasite, even if there is more than one hostOf field in the record. Table 3 and 4 show the properties of the class Parasitism, in Table 5.

[4] Available from: http://aims.fao.org/aos/agrovoc/c_3673. Access on: 24 Nov. 2019.

Table 5 – Parasitism class

Parasitism	**CClass**
Identifier	https://git.io/JT2Uk
Definition	Negative ecological interaction in which an organism, called a parasite, develops at the expense of another organism, called a host, harming it.
Comments	
Examples	Chaparral dodder in an Orange tree.

Source: Soares (2019)

Classes fulfill the function of contextualizing the organism's role in ecological interaction. The hostTo property, which in Table 4 allows to name parasitic organisms, could be used, for example, to describe the mutualistic relationship (in which both participant organisms are benefited) between a cow and beneficial bacteria living in the animal gut, since it is subordinate to a Mutualism metadata class. The hostTo predicate would be assigned to the cow metadata record and would take as values the names of these beneficial bacteria.

To exemplify how the given example of parasitism could work applying the metadata proposed in Tables 2 & 4 combined with DwC metadata, we created Schema 1.

Schema 1– Fragment of an XML record of *Telenomus podisi* (Ashmead 1893)

```
1: <?xml version='1.0'?>
2: <rdf:RDF
3: xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4: xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
5: xmlns:dwcab="https://github.com/Filipi-Soares/Agrobiodiversity-
metadata/blob/Descritores/README.md#/">
6:  <rdf:Description rdf:about="https://www.gbif.org/pt/species/1401314/">
7:   <rdf:type  rdf:resource="http://rs.tdwg.org/dwc/terms/Taxon"/>
8:    <dwc:taxon rdf:parseType="Literal">
9:     <dwc:scientificName>Telenomus podisi</dwc:scientificName>
```

```
10:     <dwc:scientificNameAuthorship>Ashmead,
1893</dwc:scientificNameAuthorship>
11:    </dwc:taxon>
12:    <dwcab:parasitism rdf:parseType="Literal">
13:      <dwcab:parasiteof>Euschistus heros Fabricius,
1794</dwcab:parasiteof>
14:    </dwcab:parasitism>
15:  </rdf:Description>
16: </rdf:RDF>
```

Source: elaborated by the authors (2020)

In Schema 1, there are three declared vocabulary namespaces: RDF namespace (line 3), Darwin Core namespace (line 4) and Darwin Core Agrobiodiversity Extension (line 5). It means that this metadata record uses metadata terms from these three vocabularies. Each namespace has an abbreviation, which reduces the length of the code (instead of repeating the full URL of the name spaces for each property, we use just the abbreviated namespace declared by the `xmlns` prefix).

This example just illustrates the work we are doing. The next step in this research is to involve the international community of practice on data management concerned about the use of biodiversity data in agriculture and food production in the metadata creation process. We believe the scientific community is going to show up the best solution for semantic problems about agrobiodiversity metadata as those shown in this paper. Thereafter, we expect the resulting metadata extension will be used by researchers all around the world.

Table 6 is a summary of the methodological principles we believe are necessary to create a metadata extension to Darwin Core. These principles are classified in a hierarchical order (e.g. 4, 4.1, 4.1.1) to highlight the relationship between them.

Table 6 – Summary of some methodological principles to create a metadata extension to DwC

1	To study and analyze the data subject that is intended to be represented with metadata, in order to define the scope and approach of the data representation.
1.1	To create a branch of classes to arrange data properties into categories.
2	To analyze the metadata of the main core of DwC to check the existence of elements for the subject data that one wants to represent with metadata.

3	To analyze the core of terms of DwC Registered Extensions[5] to verify the existence of metadata for the scope of the subject data that is intended to be represented.
4	To search for ontologies, metadata schemes, or any other pre-existing conceptual model for the subject field that is intended to be represented by metadata.
4.1	To perform a correlated analysis between the terms of the conceptual model and the DwC, if any conceptual model is found for the subject field. Then, one should classify the terms of the conceptual model into two categories: DwC equivalent terms and non-equivalent terms.
4.1.1	Use non-equivalent terms as potential new metadata.
4.1.2	Use equivalent terms as DwC metadata, from Darwin Core Task Group (2009d).
5	To adopt requirements and recommendations based on standards and data models from the scientific literature to propose the syntactic and semantic structure of the metadata extension elements, so other can use your metadata. It is suggested:
5.1	the standard ISO/IEC 11179-4 (2004), which provides recommendations with clear and exemplified definitions of best practices in creating a definition for metadata vocabulary;
5.2	the Dublin Core Abstract Model (Powell et al. 2007) as the semantic model for the "design of metadata records in terms of structural components, such as Descriptions, Statements, Properties, and (literal or non-literal) Values, in order to enable structural validation of RDF-based metadata".
5.3	The RDF Schema for modeling the metadata vocabulary (Brickley et al. 2014).
5.4	the Extensible Markup Language (XML) as the syntax for the RDF metadata application (Bray et al. 2008), but also Turtle, JSON, and other markup languages can be applied.
6	To embrace the three forms of extending metadata schemas, if applicable: a) to create new metadata elements, which should have names, labels, definitions, and functions different from the pre-existing metadata in the DwC, according to items from 6.1 to 6.2.9; b) qualifiers: a qualifier term must be bonded together with a pre-existing term in the metadata vocabulary to identify a specific value, for example, 'dc:date' and 'dc:dateRegistered', according to items from 6.3 to 6.3.2; c) encoding schemes, which provide guidelines for formatting the metadata terms and data values, as per item 6.4.
6.1	To create a terminology sheet for each term of the extension's metadata, defining its attributes such as term name, namespace, definition, and additional information that might help users to apply the metadata (see Brickley et al. 2014).
6.2	To apply the requirements of ISO/IEC (2004), from item 6.2.1 to 6.2.6, and additional

[5] Available from: https://tools.gbif.org/dwca-validator/extensions.do. Access on: 26 Aug. 2020.

	recommendations, from item 6.2.7 to 6.2.9, to create new metadata elements.
6.2.1	Write the term that represents the element in the singular form.
6.2.2	To define what the concept is in fact, without saying what it is not.
6.2.3	To set the element' definition as a very descriptive sentence, in other words, that clarifies the meaning of the term used as an element.
6.2.4	Do not use abbreviations that are difficult to understand.
6.2.5	Do not embody definitions of other metadata or underlying concepts.
6.2.6	To incorporate a controlled vocabulary to assign the terms used as metadata terms.
6.2.7	To search for a term definition in the scientific literature when the controlled vocabulary does not provide an underlying conceptual definition for the term used as a metadata element.
6.2.8	To write the metadata terms in the lower CamelCase format. For example: occurenceRemark, lifeStage, reproductiveCondition.
6.2.9	To create a namespace to identify the extension's metadata for computer systems.
6.3	To create qualifiers for pre-existing metadata in the DwC element set. Qualifiers are words that make the representation of a more specific value, i.g., 'dateAccepted' is a qualified version of the element ‘date’.
6.3.1	Do not use abbreviations that are difficult to understand in qualifiers.
6.3.2	To write the metadata qualifier terms in the lower CamelCase format. For example: nameAccordingTo, namePublishedIn, namePublishedInYear.
6.4	To incorporate encoding schemes prescribed in the DwC standard to format the data values, such as:
6.4.1	DCMIType Vocabulary: sets of classes defined by DC to describe the type of resource.
6.4.2	Dewey Decimal Classification (DDC): the set of subject classes of the CDD classification system, widely applied in libraries for classification of bibliographic documents.
6.4.3	Internet Assigned Numbers Authority (IMT): terms for media types.
6.4.4	Library of Congress Classification (LCC): the set of subject classes of the LCC classification system, widely applied in libraries classification of bibliographic documents.
6.4.5	Library of Congress Subject Headings (LCSH): the set of subject headings from the LCSH controlled vocabulary, widely used in libraries for indexing bibliographic

	documents
6.4.6	Medical Subject Headings (MESH): the set of MESH controlled vocabulary concepts, widely used for indexing medical documents.
6.4.7	National Library of Medicine Classification (NLM): the set of controlled vocabulary concepts of the NLM classification system, widely applied in libraries for the classification of medical documents.
6.4.8	Getty Thesaurus of Geographic Names (TGN): the controlled vocabulary of names of cities and other kinds of localities.
6.4.9	Universal Decimal Classification (UDC): the set of subject classes of the CDU classification system, widely applied in libraries for document classification.
7	To involve the community of practice in the construction of the metadata extension, so the data representation can meet the needs of those who will use them in practice.
8	The metadata extension archive has to be identified by a single metadata registry, including the following attributes (from 8.1 to 8.7):
8.1	definition: the subject scope of the metadata extension;
8.2	see also: a source of information about discussion groups, vocabulary or the development history of the extension;
8.3	properties: a number or enumeration value for the number of metadata elements in the extension;
8.4	name: the name given to the metadata extension;
8.5	namespace: a URI for the metadata extension (Darwin Core Task Group 2009c);
8.6	rowType: URI assigned to the term to identify the represented data class (Darwin Core Task Group 2009d);
8.7	keywords: indexing words for the extension's subject classes.
9	Each metadata element of the extension must be defined by attributes (from 9.1 to 9. 9):
9.1	termName: name of the element that can be used as a metadata field in a record;
9.2	definition: the term meaning;
9.3	see also: a source of information about discussion groups, vocabulary, or history of the term;
9.4	qualified name: the term's URI;

9.5	examples: examples of values that can be assigned to the metadata elements;
9.6	namespace: DwC terms must be identified with a URI. These URIs are grouped into collections called Darwin Core namespaces (Darwin Core Task Group 2009c);
9.7	group: the class that comprises the metadata element;
9.8	datatype: the type of value (data) that can be entered in the metadata field;
9.9	required: indicates if the use of the element is mandatory in all DwC metadata records or not.

Source: elaborated by the authors (2020)

All these principles presented in Table 6 are results of the research. Some of them were based on ISO/IEC 11179-4 (2004) (items 6.2.1 to 6.2.6 and 6.3.1), and DwC documentation (2009a, 2009b, 2009c, 2009d) (items 6.2.8 to 9.9).

The principle 4 of Table 6 recommends searching for ontologies and other metadata schemas. The recommended reference source to search for metadata schemas is the Digital Curation Centre Metadata Guidance[6] that holds a detailed description about metadata standards of five domains of knowledge: Social Science & Humanities, Physical Science, General Research Data, Earth Science, Biology. To search for ontologies for the biodiversity field, some recommended reference sources are Planteome[7] — for plant ontologies —, Ontobee — for ontologies related to biodiversity, agronomy and many other scientific fields —, and Open Biological and Biomedical Ontology (OBO) Foundry.

We expect these principles to be useful for other researchers to create their own metadata extensions to DwC.

6 Conclusion

It is still necessary to discuss what would be the best way to relate the metadata extension to the core of DwC terms. We know the principles presented here are broader, and as so, require

[6] Available from: https://www.dcc.ac.uk/guidance/standards/metadata. Access on: 22 Out. 2020.

[7] Available from: http://browser.planteome.org/amigo. Access on: 22 Out. 2020.

some refinement and organization as a more robust methodology. This is going to come along with the participation of the community of practice on the creation of this metadata extension.

We believe the principles presented in this paper can contribute to giving the guidelines for metadata creation to communities concerned with agrobiodiversity data management and even other communities using DwC or other metadata standards. There is a long way ahead to develop a complete methodology for metadata extension creation that can be applied to any context related to agriculture, biodiversity, and food production.

References

Alves, R. C. *Web Semântica: uma análise focada no uso de metadados*, 2005. Universidade Estadual Paulista, PhD dissertation.

Alves, R. C. "Metadados para representação e recuperação da informação em ambiente Web". *Proceedings of the 4th Seminar on Museum Information Services*: Digital Information as Cultural Heritage, São Paulo, 2016.

Arnaud, E., et al. *Final report of the task group on GBIF data fitness for use in agrobiodiversity*. GBIF, 2016.

Body, G., et al. "Applying the Darwin core standard to the monitoring of wildlife species, their management and estimated records". *EFSA Supporting Publications* vol. 17, no. 4, 2020, p. 01-77, doi: 10.2903/sp.efsa.2020.en-1841. Accessed 3 march 2020.

Bray, T., et al. *Extensible Markup Language (XML) 1.0*. W3C, 26 November 2008. https://www.w3.org/TR/REC-xml/#sec-intro. Accessed 25 November 2020.

Brazilian Agricultural Research Corporation (EMBRAPA). *Embrapa em números*. Brasília, 2019.

Brickley, D., et al. *RDF Schema 1.1: W3C Recommendation 25 February 2014*. W3C, 25 Feb. 2014. https://www.w3.org/TR/rdf-schema/#bib-RDF11-CONCEPTS. Accessed 25 November 2020.

Cain, M. L., et al. *Ecology*. 4th ed. Sinauer Associates, 2017.

Cassini, S. T. *Ecologia:* conceitos fundamentais. PPGEA-UFES, 2005.

Chu, H. *Information representation and retrieval:* an overview. Information Today, 2005.

Convention on Biological Diversity (CBD). COP 5 *Decisions: Agricultural biological diversity: Review of phase I of the programme of work and adoption of a multi-year work programme: Annex: Programme of work on agricultural biodiversity*. CBD, 2000. https://www.cbd.int/doc/decisions/COP-05-dec-en.pdf. Accessed 3 march 2020.

Coyle, K. DCAM Explained, 2008. http://kcoyle.net/dcam_simple.html. Accessed 3 march 2020.

Darwin Core and RDF/OWL Task Groups. *Darwin Core RDF guide*. TDWG, 2015. http://rs.tdwg.org/dwc/terms/guides/rdf/. Accessed 3 march 2020.

Darwin Core Task Group. *Darwin Core text guide*. TDWG, 2009a. http://rs.tdwg.org/dwc/terms/guides/text/. Accessed 3 march 2020.

Darwin Core Task Group. Simple Darwin Core. TDWG, 2009b. http://rs.tdwg.org/dwc/terms/simple/. Accessed 3 march 2020.

Darwin Core Task Group. Darwin Core XML guide. TDWG, 2009c. http://rs.tdwg.org/dwc/terms/guides/xml/. Accessed 3 march 2020.

Darwin Core Task Group. *Darwin Core basic vocabulary.* TDWG, 2009d. http://rs.tdwg.org/dwc/. Accessed 3 march 2020.

DCMI Usage Board. *DCMI Metadata Terms*: date, 2020. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/terms/date. Accessed 3 march 2020.

FAO. *Why bees matter: the importance of bees and other pollinators for food and agriculture*. FAO, 2018.

ISO, and IEC. *ISO/IEC 11179-4: information technology: metadata registries (MDR): part 4: formulation of data definitions*. 2nd ed. ISO, 2004.

ISO, and IEC. *ISO/IEC 11179-6: information technology: metadata registries (MDR): part 6: registration*. ISO, 2015.

Kenis, M., et al. *Guide to the classical biological control of insect pests in planted and natural forests*. FAO, 2019.

Lourenço, C. A. "Representação de informação, metadados, interoperabilidade e recuperação da informação na atualidade". *Proceedings of the 2nd Seminário do Grupo de Pesquisa <MHTX>,* Belo Horizonte, 2016.

National Information Standards Organization. *Understanding Metadata*. NISO Press, 2004.

National Information Standards Organization. *A framework of guidance for building good digital collections*. 3rd ed. NISO Press, 2007.

Odum, E. P., and Barret, G. W. *Fundamentals of ecology*. 5th ed. Thomson Brooks/Cole, 2005.

Oxford University Press. Oxford Learner's Dictionaries, 2020. https://www.oxfordlearnersdictionaries.com/us/definition/english/date_1?q=date. Accessed 3 march 2020.

Pacheco, D. J., and Corrêa-Ferreira, B. S. “Parasitismo de Telenomus podisi Ashmead (Hymenoptera: Scelionidae) em populações de percevejos pragas da soja”. *Anais da Sociedade Entomológica do Brasil* vol. 29, no. 2, 2000, p. 295-302.

Pinto-Coelho, R. M. *Fundamentos em ecologia*. Artmed, 2007.

Pomerantz, J. *Metadata*. The MIT Press, 2015.

Powell, A., et al. “DCMI Abstract Model”. Dublin Core Metadata Initiative, 2007. http://dublincore.org/documents/abstract-model/. Accessed 3 march 2020.

RDF Working Group. “Resource Description Framework (RDF)”. *W3C Semantic Web*, 25 Feb. 2014. https://www.w3.org/RDF/. Accessed 3 March 2020.

Riley, J. *Understanding Metadata: what is Metadata, and what is it for?: a primer*. NISO, 2017.

Soares, F. M. Princípios para a criação de uma extensão de metadados sobre interações ecológicas na agrobiodiversidade para o padrão Darwin Core, 2019. Universidade Federal de Minas Gerais, Master’s thesis. http://hdl.handle.net/1843/33387. Accessed 3 march 2020.

Soares, F. M., et al. “Linking Agrobiodiversity Data through Metadata Standards”. *Biodiversity Information Science and Standards,* vol. 4, 2020, p. e58928. doi:10.3897/biss.4.58928. Accessed 3 march 2020.

Soares, F. M., et al. “Darwin Core for Agricultural Biodiversity: A metadata extension proposal”. *Biodiversity Information Science and Standards* vol. 3, 2019, p. e37053. doi: 10.3897/biss.3.37053. Accessed 3 march 2020.

Sorci, G., and Garnier, S. “Parasitism*”. Encyclopedia of Ecology*, edited by S. E. Jørgensen and B. F. Fath. Elsevier, 2008, pp 2645-2650. doi: 10.1016/B978-008045405-4.00814-4. Accessed 3 march 2020.

Stein, R. T. *Ecologia geral*. SAGAH, 2018.

Wieczorek, J., et al. “Darwin Core: an evolving community-developed biodiversity data standard”. *PloS one,* vol. 7, no. 1, 2012, p. e29715. doi: 10.1371/journal.pone.0029715. Accessed 3 march 2020.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. “The FAIR Guiding Principles for scientific data management and stewardship*”. Scientific data*, vol. 3, 2016, pp. 160018. doi: 10.1038/sdata.2016.18. Accessed 3 march 2020.

Zeng, M. L. “Metadata basics”. Version 2.0, 2015. https://marciazeng.slis.kent.edu/metadatabasics/types.htm. Accessed 3 march 2020.

Zeng, M. L., and Qin, J. *Metadata*. Neal-Scguman Publishers, 2008.

Recived: 26/08/2020 Acepted: 21/11/2020