Topic 1: Metadata

If you can't find a standard…
Try building up your own one!

"The nice thing about standards is that there are so many of them to choose from."

Admiral Grace Hopper

Diversity of Biological Data

If no metadata schema aligns with the user requirements (and the reader is convinced that there is none!), it may be necessary to design a standard that covers the technical necessities of that specific project. However, adapting preexisting schemes, or type definitions into the new requirements may also be a good option. Defining a metadata schema from scratch is a complex and arduous task.

The very first step is to understand the requirements of the data that needs to be described in the moment and at a certain point of time. Questions such as the future loading of the data into the web, or the possibility of sharing the data with other people is rellevant to select a metadata schema or another. Additionally, it is also crucial to understand the data itself. What is considered to be the information object in our data? Which characteristics are crucial, interesting or irrelevant for

describing this object? Depending on the answers to these questions, many possibilities arise:

1. The structure of the metadata scheme will allow the user to understand better what the metadata is trying to describe. For example, the metadata record can be

   composed of singular metadata elements or by groups of metadata elements, like in a cooking recipe, where we find information within the ingredients, and information within the cooking process. Establishing the structure of the metadata *a priori* is crucial to allow the interoperability of the data afterward.

DESCRIBING, DESCRIBING, DESCRIBING

A new metadata schema often needs to be understood by others besides the creator, and this requires an understanding of the different elements and possibilities within the metadata schema. An additional document defining your metadata record will also be necessary to read your metadata. This is called **Document Type Description (DTD)**, and its purpose is to define each element that could be included in the metadata, and the characteristics of it (controlled vocabulary, length of the element, etc.). In the case of the internet, but also when a human being reads through metadata, it is important to find these "instructions for use" easily. The DTD can be at the beginning of your metadata document, or live somewhere in the web, in which case an

2. If the project includes dissemination and sharing beyond the initial institution, it will probably require uploading the data to the web, and therefore the metadata should be built according to an established scheme, even at a simple level, readable for internet browsers. At this point, if already existing schemas have been discarded, it may be a good idea to opt for the simple metadata model "**Resource Description Framework" (RDF)**, which defines three categories in any definition statement: Object (described object), Subject (specific characteristic) and Predicate (field at which the characteristic applies).The RDF is a simple general method for conceptual description in web resources which basically consists in connecting the three definition categories in triples of information. (Figure 2). If your metadata record has to live in the internet, a markup language (XML, HTML, SGML or GML) will also be required for the browser to understand the record.
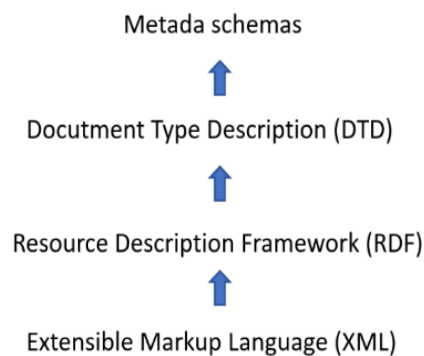

Figure 1. Hierarchical representation example of the dimensions of a metadata schema. From the very language to the final schema.
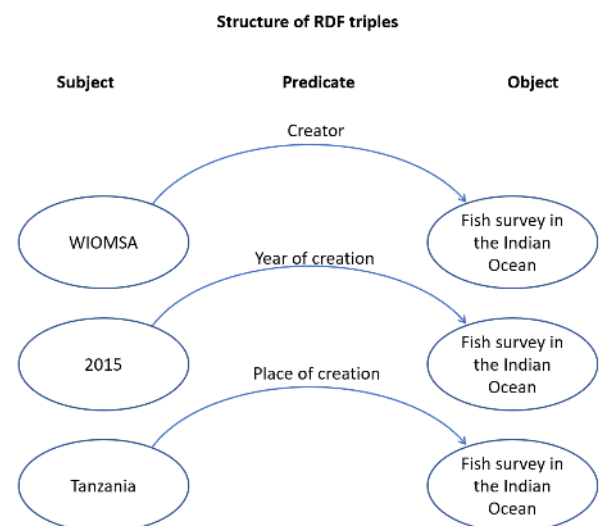

Figure 2. Triple of information example applied to Fish survey data in the Indian Ocean.

3. When deciding the elements that your metadata will have (i.e. when writing the DTD) a good idea may be to include useful elements from other schemas (i.e. creator, type of data, species, location). However, make sure that your elements do not overlap!

4. Try to minimize the elements. Instead, create sub-elements of already existing elements (i.e. better to create an element like "organism" with sub elements, than 6 elements for Kingdom, Class, Order, Family, Genus and Species).

5. Define which vocabulary best fits your requirements for each defined element. Using a controlled vocabulary will reduce typing errors and will facilitate interoperability, but it will also limit the descriptive power of the elements. No coherence is required between elements: an element may be improved with a controlled vocabulary (i.e. "month of the year "), while others may not (i.e. "Favourite sentence in English language").

REFERENCES
1. 1. Baca, Murtha (editor). 2016. Introduction to Metadata. Third Edition. Los Angeles: Getty Research Institute. Online Edition: http://www.getty.edu/publications/intrometadata
2. Steven J.Miller. "Metadata Resources" University of Wisconsin-Milwaukee. Last update in June 1, 2011. Some revisions in February 2020.
3. National Information Standards Organization (U.S.). Understanding Metadata. Bethesda, MD: NISO Press, 2004.

ABOUT THIS POLICY BRIEF

This Policy Brief is part of a series aiming to inform policy-makers on the key results of the NeDiT research project and provide recommendations to policy-makers. The series of NeDiT Policy Briefs can be found at http://nedit.net/downloads/. This publication was commissioned, supervised and produced by NeDiT project partners.

DISCLAIMER

The policy recommendations made do not necessarily reflect the views of the ZMT, IMS or its

IMPRINT

Authors:

Cescc Gordo[1], Hauke Kegler[1], Daudi Msangameno[2], Christopher Muhando[2] and Hauke Reuter[1]

The authors work at, or are affiliated with, [1]Leibniz Center for Tropical Marine Ecology (ZMT) and [2]Institute of Marine Sciences Zanzibar (IMS).

You can find more information about the project here.

The NeDiT project was funded as part of the MeerWissen Initiative by GIZ, you can find