The Meaning of Metadata

Ilias-Vasileios Varelas
Department of Informatics
Ionian University
Corfu, Ionian Islands, Greece
p18vare@ionio.gr

ABSTRACT

In the following paper we will explain the meaning of the term *Metadata* and its different applications as well as difficulties. Also, emphasis is given in the category of Web metadata for the use of locating a web resource.

In this work, there are chapters with information relevant to the understanding of the need for the creation of metadata, as well as information about the troubles that academic members, publishers, librarians and also volunteering experts had to go through in order to create the form of metadata that we use today in our cataloging procedures. Other chapters also refer to the world wide web and its influence on metadata applications and usage, cataloging difficulties that exist due to the size of the world wide web but also information about the human-mediated directories and search engines.

There are also relevant information about the way that each and every method of cataloging and indexing on the web works, as well as some data for the most used search engines in 2021.

KEYWORDS

Metadata, Meaning of Metadata, www, World Wide Web, crawlers, robots, Web Directories, Catalogues, XML, Identifiers, Dublin Core, Dublin Core Metadata Initiative, DCMI

1 Introduction

Let's start with an example. In today's world every file in physical format is being digitized and stored in a computer, either locally on our machines, or online in some cloud storage service or web server. Given the importance of each one of those files, we need to have a system that allows us to find and access our files at any given moment. For the files that are located on our local machine, it's relatively easy, because the computer has been using catalogues internally in order to keep track of every file, code or executable it needs to keep working.

The difference between the locally saved files and files stored online is based on the size of the World Wide Web (WWW). In order for us to be able to find and recover a specific file in the chaos of information that the WWW provides, we need a system for reaching every web page that exists and cataloging every result of the process. This process is called Crawling.

For the Crawling process to become effective, we need to add some key information on every web page that we want to be able to easily locate and access later on. This information is called Metadata. In many published papers, Metadata is defined as “Data about Data”, which is a simple but fairly accurate definition of the term.

2 Basic Terminology

The term metadata is most commonly described as data about data or information about information. Metadata exist in both print and electronic forms of items or files. There are some classifications that can be applied on the metadata based on their purpose, such as technical metadata for technical processes, rights metadata for rights resolution, preservation metadata for digital archiving and descriptive metadata for the description of an item’s content, to name a few.

Depending on the application of the metadata, there are some variations as to how much detail is stored in them.

3 Size of the World Wide Web

The size of the World Wide Web is increasing by the second[1]. Back in April of 2000, Netcraft carried out a survey that pinged almost 14.5 million domain names. Today that number has climbed to surpass the 1 billion mark for total domain names that are either active or inactive in the WWW.[3]

Keeping in mind the extreme size of this very big catalogue that we call the WWW, it should be pointed out that before the metadata were created, there was no other way to filter through all the available files or items that are stored online in order to pinpoint a specific one. That's because neither the Internet nor the WWW were designed with the cataloging of their contents in mind. The TCP/IP Protocols are created solely for package transportation from the sender’s IP address to the recipient’s, whereas the Hyper Text Transfer Protocol only deals with the delivery of hyperlinked information.

With the aforementioned in mind, the creation of modern tools that took advantage of the metadata began. The two most common tools for locating a web resource are the Directories and Search Engines.

4 Directories, Search Engines and Crawling

Directories are listings of network resources created by real people, who oversee, catalog and classify web resources that they deem fit based on factors such as accuracy, authority and currency. The scope of a directory is also variable. It can be broad or specific and the content is available either through searching or by browsing a hierarchical set of subject headings. There are a few examples to be given, such as the World Wide Web Virtual Library and Yahoo! that are directories with a wider span of results, but there are also examples like the Art, Design, Architecture & Media Information Gateway (ADAM) that have a more subject-focused result span.

Search Engines, also known as “spiders”, “crawlers” and “robots” are automated algorithmic systems that endlessly travel through the Web visiting any link they find available, while at the same time they are cataloging every page they encounter, thus creating a massive directory, or catalog, of indexed pages. The scale of the indexation depends on the site itself and the way that it is built. Search engines typically return a vast amount of results based on our search, they provide powerful searching queries with easy to use graphical interfaces and cutting edge algorithms that can “smartly” recognize the meaning behind each search that is taking place.[4]

There are a number of World Wide Web Search Engines available in today's internet, but the vast majority of holdings belong to Google Search. For reference some popular search engines are :

1. Ask.com
2. DuckDuckGo
3. Microsoft Bing
4. Yahoo! Search
5. AOL

Google’s algorithm is proprietary and is not available to the public. Although, the crawling process that it utilizes is explained in numerous sites on the Web. First, it begins crawling a list of web addresses and sitemaps that have been provided by the website owners. While the crawler is visiting a specific site, it discovers and catalogues the links that exist in the site while at the same time it uses those links to visit and discover new sites. The algorithm pays special attention to new sites, changes that may have taken place to already catalogued sites and dead-end links. Every link that is recorded from the crawler is being stored in Google’s servers for it to be able to get included in Google Search’s database of potential results in a search query. The last functionality of Google’s crawling algorithm is the organization of information based on the indexing factor. Whenever a site is being rendered by a robot, Google takes notes of some key signals that exist on the page such as keywords, HTML meta tags and website freshness, to name a few. All that information is stored in the Search index of the database.

Google is only one of the many present tools that we have in our armory in order to manage Web resources. Back in the 1990’s, there were only a handful of tools, with much less capabilities than they have today and those were only created after Tim Berners-Lee invented the Web in 1989. Soon after their impressive invention, they went on to create one of the first ever web browsers, the World Wide Web Virtual Library, that is maintained and fully working up until today. The WWW Virtual Library is the oldest catalogue of the Web[5]. Even today, the Virtual Library is run by volunteers. Each and every one of them is responsible for the content of their own pages, with a few general guidelines applying. Each volunteer is only maintaining pages in particular areas in which he has an expertise.

5 Cataloging Difficulties

Even after all of the evolution that the WWW has undergone through all the years, there are still a few problems and difficulties that have remained unoptimized. Human-mediated directories generally provide high search precision at the broad subject level, and are normally considered to provide higher-quality information overall because of the human intervention and expertise in the indexing and classification process. However, this intervention is expensive and time consuming when it’s compared to algorithmic indexing. Also, the labor that is required, is not sufficiently scalable to provide comprehensive up-to-date coverage of the Web.

Another problem with the human-mediated method of cataloging is the granularity of the resources that need to be described; should there be a unified description of the whole website or should there be a description for each individual page of the website? There hides a tradeoff that will always need to be made.

The spider-based search engines also suffer from a number of serious problems that affect their ability to create and provide an index that is both comprehensive and up-to-date. Such problems occur from pages that provide dynamically created information from databases based on user input. This type of information is referred to as “the Hidden Web” because it outweighs the indexing capabilities of the Web crawlers. Given the automated nature of the Web crawlers, cataloging large automatically-indexed databases often results in extremely large results sets, which are frequently unusable despite increasingly sophisticated and state-of-the-art information retrieval tools or artificial intelligence sorting algorithms.

In a study carried out by Steve Lawrence & C. Lee Giles from the NEC Research Center back in february of 1999 the results showed that the search engines are having a hard time indexing the entirety of the Web. The study discovered that the combined coverage of the 11 search engines used, only covered about 42% of the total number of unique indexable pages on the Web. No individual search engine managed to index more than 16% of the sites available[6]. After the publication of this study, the search engines started quoting ever-lager numbers of pages crawled, in order to compensate for the embarrassing exposure.

Even if it looks like both the methods are suffering from distinct and different type of problems, there is common ground when it comes to the source of all the aforementioned problems, and that's the ambition guiding the actions. The sad truth is that the Web is just too big to be catalogued or indexed by one single organization or service even if there are people, algorithms or both, working on it.

The solution to this problem, if there is any, is probably hidden behind the distributed metadata catalog model, which is the model used by the WWW Virtual Library for indexing purposes, although the efforts of the volunteers on it have been proven insufficient to keep up with the growth of the Web.

6 Metadata Applications

The idea is that the information structure and content of Web metadata should capture the essence of the Web's resources as well as describe and facilitate the various tasks for which they were created. Sadly, due to the size and complexion of the Web, there is an extensive collection of objects to be described. This points to the fact that the applications of metadata can only be limited by the imagination.

To help with understanding the complexity and structure of a metadata record, we will take a look at the metadata standard called Dublin Core. The Dublin Core Metadata Initiative (DCMI) started back in 1995 as a joint project among professionals from the publishing, library and academic communities. One outcome of this effort was the Dublin Core Metadata Element Set, which became a NISO standard in 2001 and an international standard (ISO 15836) in 2003. The DCMI standard includes fifteen metadata elements for describing information resources. These are : *title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage* and *rights*. Some of these elements refer to the content of the resource, other to the intellectual property and some of them to a particular version of the item.

The most common application of web metadata is generally referred to as *resource discovery* because it's role is to assist a Web user in his effort to pinpoint a specific Web resource. The existence and carefully created metadata descriptions can lead to much greater search precision and more accurate relevance ranking of the results that are retrieved by the search engines.

Even after their usage as a filter mechanism, to be put vaguely, metadata is still used in order to provide more information to the user about the results that are available to him. Short descriptions follow each search result, printing just the necessary amount of information in order to give a general idea of what that website's content is.

Metadata is also often used in the management and administration of digital networked resources. This type is referred to as *administrative metadata*, and is essential to ensuring that web resources are kept up to date or provide any relevant information about rights restrictions that may prohibit their distribution over the Internet.

Metadata also play a big role in publishing and communication cycles. The practice that has given metadata such a supportive role in the publishing communities is standardization. Standardization is the process of building consensus around already proven practices with acceptable results, in the formatting and use of metadata for specific applications, so that machines can interpret and exchange this information efficiently[2].

Standards-building is a process in which book and journal publishers should invest and participate. Despite the fact that a much greater proportion of journal content than book content is digitized, publisher-driven standardization initiatives in book publishing are more advanced than in journal publishing. Book publishers have been using standardization for capitalizing reasons, aggregated bookselling, firstly via wholesalers and later through the Internet, which has forced them to rely on standards from supplying promotional metadata.

The fact that there are so many possible implementations for metadata is a blessing and a curse. The large variety of scenarios can make the standardization really hard, so questions are starting to appear. How much detail about a Web resource should a metadata record contain? How many catalog records should be created for a given Web resource?

Since the 1950's, the information retrieval tools have come a long way in creating a dynamic list of results based on a user's search. At first the idea of an Information Retrieval (IR) tool arose in 1945 from Vannevar Bush, and during the years 1955 to late 1960's, there were a lot of experimentation and seminars regarding IR tools. It was during the 1970's that the IR started to mature into real systems. The 2 decades that followed were the beginning of the Databases era, where databases were introduced and implemented as an information tool.[7] Today, information retrieval tools have gotten even better, with the ability to recognize an image's content and retrieve information about the original image, or fetch files that contain one specific word that the user has searched for.

This evolution has led to the increase of user expectations regarding IR tool's capabilities and combined with the flexibility and diversity of the hypertext information environment, they render the Web-Bibliographic cataloging analogy partially invalid.

Another significant difficulty arises from the need to describe relationships between networked resources and other objects; What should metadata describe?

By definition, metadata should describe information about an object that is itself data, like a web page, a digital image or a database. For networked resources however, this method isn't necessarily useful. For example, if a user is looking for information about a painting, or an image of one, he will perform a search based on the original properties of the item such as, Creator: Picasso, Date: 1937, instead of, Creator: Art Scanning Company, Date:2000.

7 What Isn’t Metadata

Metadata is a term that usually refers to standardized and structured information that computers can interpret and use. The boundaries of this definition often overlap, yet are not to be confused with two related terms, Extensive Markup Language *(XML)* and *Identifiers*.

XML is a language for presenting and expressing rules that give structure to any kind of information, including metadata. XML has been widely adopted because it was designed for precisely the kind of data transfer that comprehensive electronic publishing requires. It also provides an application-independent method for sharing data, and because it is free to license, XML can save publishers money through the use of inexpensive, off-the-shelf tools. A large part of its power comes from the nearly universal support it receives from product vendors, standards bodies, academia, and the open source community

XML uses a simple syntax that is easy for both humans and machines to understand. It consists of tags that surround information, like <head> … </head>. These tags can also be associated with attributes, also known as name-value pairs.

Additionally, XML provides a document that acts as a description of every available element that is supported by the language. This document is called Document Type Definition (DTD). It includes the available types of elements, the order that they should be in and how they interrelate. DTDs can vary in use and rights of access. Some are proprietary and are being used for a company's internal use only, while others are standardized and freely available. DTDs can also be created freely, which is a feature of the XML that makes it a great tool for structure and presentation of every type of information.

An XML schema, also called XSD file, is itself an XML document that enhances developers capabilities for document and information validation as well as adding more tools to their toolkits for creating their own XML-based formats. An XML schema is, to be put bluntly, a more advanced version of a DTD. Whereas DTDs only allow for relatively simple data types creation and management, a schema has a set of powerful, flexible semantics for defining what an XML file can contain.

On the other hand, metadata is also being mistaken for identifiers.

Identifiers are names or strings adhering to certain conventions that, if properly implemented, guarantee uniqueness. While standard identifiers have been in use for decades, the ability to easily recognize electronic content has become extremely important in the electronic publishing and e-commerce platforms. Even though identifiers and metadata are not the same term, their combination turns out to be really useful.

8 Why is Metadata Important

Metadata turned out to be an extremely useful tool for professionals but also for everyday users that are looking either to create and categorize information or they are simply searching for a new recipe online.

Carefully crafted metadata has become one of the tools that information professionals are using to exploit and optimize the cataloging process.[8]

One of many advantages that the metadata provides is the enhanced effectiveness of searching through the existence of rich and carefully crafted descriptive metadata. Descriptive metadata can also link various repositories and items that do not belong in the same categories, as long as their descriptive metadata records match.[8]

Another advantage is the context retention that they can provide. In museum, archival or library repositories, metadata can serve as a simplifying mechanism for the complex collections of objects that exist. They play an important role in documenting and maintaining important relationships, as well as indicating the authenticity, structural and procedural integrity and degree of completeness of information objects. In an archive, for example, by documenting the content, content and structure of the record, metadata can be the determining factor between the authentic record and some decontextualized information.[8]

Conclusions

Metadata is a powerful tool that can help every user, professional or not, locate, recognize, manipulate and categorize any type of file, both in physical or electronic forms. The fields in which metadata can be implemented are endless, but the most common is the Web resource localization. The advantages that they have to offer are numerous and several researches show that the science of Information has barely cracked the surface. The teaching and correct usage of metadata can only optimize our everyday procedures.

ACKNOWLEDGMENTS

This paper is written in the context of a class of my university addendum. This is not a scientific or published paper.

REFERENCES

[1] Gill, Tony. "Metadata and the World Wide Web." Introduction to Metadata: Pathways to Digital Information. Version 2.0. Edited by Murtha Baca. (J. Paul Getty Trust, 1998)

[2] Amy Brand, Frank Daly Barbara Meyers ,Metadata Demystified: A guide for publishers, http://www.niso.org/publications/metadata-demystified-guide-publishers, NISO Press, July 2003

[3] Research from the University of Tilburg, September 2021

[4] How crawling works. https://www.google.com/search/howsearchworks/crawling-indexing/

[5] WWW Virtual Library : http://vlib.org/admin/AboutVL

[6] Steve Lawrence,C. Lee Giles,Accessibility of information on the World Wide Web, https://www.researchgate.net/publication/220097958_Accessibility_of_information_on_the_World_Wide_Web, February 1999.

[7] Michael Lesk, The Seven Ages of Information Retrieval, International Federation of Library Associations and Institutions,https://archive.ifla.org/VI/5/op/udtop5/udt-op5.pdf , March 1996
[8] Anne J. Gilliland, Setting the Stage, Paul Getty Trust, http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf , 2008