# MAKING YOUR INFORMATION AUTOMATIC

**Vicent Gil Esteve**: Fourth year student of the Information and Documentation degree at Universitat de Barcelona, vicentgilesteve@gmail.com

**Albert Rubio Velasco**: Fourth year student of the Information and Documentation degree at Universitat de Barcelona, albert.rubio.velasco@gmail.com

**Short summary:** *The social network phenomenon has led to a new generation of Internet users with a slew of new habits and customs in the realm of information usage. Web users want to consult a variety of information sources at high speeds in a world of constant change. In this paper we review the various tools, platforms, and schemes that allow us to be more efficient in describing, using, and sharing information. We will talk about the standard of the 3 Rs: reduce, reuse, recycle, and how it applies to information management. Efficiency is reached with automation, which is possible through two methods: importing (using information from many sources in a single place) or exporting (content is entered only once, and is published or shared at many platforms automatically). We will see how metadata can be re-used by some tools automatically, or how some tools like "IFTTT" (If This Then That) can help us with "non-librarian-described" documents and formats.*

**Keywords:** information sustainability, OAI-PMH, interoperability, IFTTT, information reuse, metadata

## Introduction

Nowadays we face a proliferation of daily information consumption, we continually consume it: blogs, social networks, instant messaging, media etc.

All around us, hardware is becoming more and more powerful as it is able to process more and more bytes. This improvement in hardware makes sense because every day billions of new documents are created, however, is this sustainable? How can we be more effective as managers, consumers and creators of information? Can we make different platforms understand and communicate with each other? Can we reuse information that has already been created?

## Background

Since the emergence of the Internet, there has been a boom of new ways to consume, produce, and share information. Early Internet platforms were part of what we call Web 1.0, which was when static web pages were created by webmasters, which did not allow for user input. For example, users could not comment, like, share, or review the pages they came across on the Internet. In present day, the web is now characterized by high user participation and involvement.

One of the problems that emerges from this increased user participation is that the user does not know how to describe and manage information. Therefore the average user does not create professional descriptions of the documents they upload and share. Even if the user documents their information, their descriptions are usually not accurate and limited to a title and few keywords. Most non-librarian users do not use metadata, indexing, classifications or controlled vocabularies. This is understandable, since an accurate description of a document, as librarians understand it, usually involves a lot of time. Besides this, duplication also appears as another consequence of the average user as a creator of content. In short, the Internet is a mass of poorly described and duplicated content.

The technological development of recent years and the development of web 2.0 is what allows us to use many web tools to manage the constantly growing information.

Junichiro Koizumi told a G8 summit in 2004 what would become a standard throughout information management, and not just in the context of a library: Reduce, Reuse and Recycle.

From the above, we can show and define that we face two types of online information:

- Information structured according to library standards: information created with the intention of making it recoverable in the future (in a catalog, database, or repository). These descriptions are created with previously fixed standards. In general, this type of information is more serious and formal.
- Information structured without library standards: Here we refer to Tweets, YouTube videos, messages on social networks, forums posts, and many other similar pieces of information. All of this information is created by users who ignore how to do an accurate description of documents and items.

By reusing the information available on the web, we save time, money, and effort. This is why the reuse of information should be a priority for us, as information managers, if we want to be truly sustainable and efficient.

Currently, information science conducts data management through various standards. This standardization facilitates interoperability between different entities and platforms. Interoperability is only possible if a single document can be interpreted by different machines, software, or people; therefore, it is necessary to describe and identify each part of it univocally. New standards have also appeared that are responsible for facilitating this interoperability. These standards facilitate the exchange of data and usually have a flexible structure (normally with repeatable elements and without essential ones).

## Goals

Our main objective is to review some available options that will allow information managers to increase efficiency through reusing information. Here are some specifics:

- Exemplify how metadata can be extracted from web elements or even library catalogs.
- Explain how this metadata can be converted to the standard that suits us in each case.
- Speak of good practices that will prevent duplicates on the Internet
- Explain how the use of "IFTTT" can help us save time when used in social networks, blogs, etc...

## Methodology

Our investigation began after we came accross an academic work that introduced us to the topic. The work itself was asking about the results of automized information retrieval and sharing. We used and combined free Internet tools with our knowledge of metadata to map data and transform that information into more interoperable languages.

The academic work outlined above was the basis of this paper, and go reached this paper by reading literature. Also, we interviewed professionals in the management and automation of metadata industry in order to have a more global view obtained from primary and secondary sources of information.

## Results

### How to extract metadata from web content

Every web page should have certain metadata in the header to ensure that search engines can provide and retrieve important information. Despite the importance of that metadata standards, descriptors currently used in most web pages are based on requirements from technological giants like Facebook and Google. In almost every website Open Graph can find for example Open Graph metadata and the necessary tags that allow sites like Google or Facebook to index and therefore read our site correctly. Open Graph is based on Dublin Core, here we can see the similarities:

```
<meta property="fb:admins" content="227104224097414"/>
<meta property="og:url" content="http://www.robotverd.cat"/>
<meta property="og:title" content="Robot Verd"/>
<meta property="og:site_name" content="Robot Verd"/>
<meta property="og:description" content="El teu lloc web d&#039;Android en
Català. Notícies, smartphones, apps, jocs, root..."/>
<meta property="og:type" content="website"/>
<meta property="og:image" content="http://www.robotverd.cat/wp-
content/uploads/2013/10/faceicon.png"/>
<meta property="og:locale" content="ca"/>
```

Figure 1: Open Graph meta tags

Our job as information managers is to work with structured information using open standards such as Dublin Core, and to accomplish this task we use an array of online tools. A good example is *DCdot Tool* that allows us to recover the Dublin Core metadata from a web page. If the URL we introduce in DCdot has no DC metadata associated, the tool will search any metadata on the page within the header and will return the results in DC format. This is an example of real-time mapping of metadata and conversion to DC.

If the extracted metadata does not match the format that best suits our needs, we can use XSL style sheets for mapping between different standards. Thus, for example, we can convert our MARC record metadata to Dublin Core 21.

### Making open and sustainable metadata: OAI-PMH

OAI was established with the mission of developing and promoting interoperability standards to facilitate the efficient dissemination of content on the Internet (Barrueco & Coll, 2003). Openly, the consultation platform (which in the literature is called service points) may be in contact with partner repositories (data providers). In other words, OAI-PMH makes content retrieval easy between repositories that join this initiative.

The language of OAI-PMH allows us to make 5 types of queries to retrieve records, identify the repository and make it individually or in a list (GetRecord, Identify, ListIdentifiers, ListRecords, ListSets or ListMetadataFormats). A good way to experiment with these types of queries can be to use the Repository Explorer of the Vernmon University (http://re.cs.uct.ac.za/).

To make the data interoperable, repositories have to encode information in unqualified Dublin Core, which means without using attributes (Weibel, 2007). Moreover, the elements must be structured with XML.

```
<dc:title xml:lang="en">Grassmann's space analysis</dc:title>
<dc:creator>Hyde, E. W. (Edward Wyllys)</dc:creator>
<dc:subject>LCSH:Ausdehnungslehre; LCCN QA205.H99</dc:subject>
<dc:publisher>J. Wiley &amp; Sons</dc:publisher>
<dc:date>Created: 1906; Available: 1991</dc:date>
<dc:type>text</dc:type>
<dc:identifier>http://resolver.library.cornell.edu/math/1796949
   </dc:identifier>
<dc:language>english</dc:language>
<dc:rights xml:lang="en">Public Domain</dc:rights>
```

Figure 2: Example of Dublin Core structure provided for OAI.

In short, if we save a document and its description is in an OAI-PMH repository, we are ensuring that it can be reused by other users, and also its description can be imported by other institutions by extracting metadata.

While OAI-PMH is an open, transparent, and flexible way for those who want to leave their data in open repositories, it has drawbacks: for example, the need to create the repository, position it on the web, share it with other institutions, etc. However there are emerging alternatives to the OAI-PMH protocol, such as micro-data standards by “schema.org” or models of semantic web such as the RDA, or RDF (metadata embedded in a HTML5 code).

**How to reuse information in social networks**

If This Then That is one of many web tools that allows you to create automated processes between social networks and various services. The main downside of this choice is the fact that IFTTT is a privative initiative and the internal process is invisible to the user. But all the features of IFTTT help us achieve our goals more efficiently. IFTTT works with different API's of web services, so it becomes the intermediary between our Vimeo, Wordpress, Facebook, Twitter or any other accounts. We must emphasize that this tool presents endless possibilities.

In our case, we have worked with Vimeo, but IFTTT is able to work with data from more than 50 different platforms. IFTTT can help us share the same content in many social networks at the same time, or to collect information from other platforms.

**How to make a more sustainable Internet**

Below is a summary of best practices that would provide a more sustainable and effective Internet:

- Describe content correctly using previously mentioned tools and standards.
- Avoid duplicate content. If necessary, we can implement the use of *nonfollow* tags in the web site’s code, or within the link code that prevents it from being indexed by a search engine.

- Do not copy (legal or illegal) content from other websites, do a direct link to the original source content.

These practices should be delegated by large companies such as Google, Yahoo, and Microsoft, among others. By doing this, the habits of users could change for the better and thus improve the quality of Internet.

## Discussion and conclusions

After the study, we have affirmed that the reuse of information is a very possible reality with today's technology and software. The work needed to handle such information management is labor intensive (metadata extraction, application of style shifts, imports, repository creation), however, this can be reduced using the proper software.

We as information management professionals are well aware of the need for structuring and the standardization of information for proper preservation, so it is imperative to use these standards if we want to facilitate a lasting and efficient use of information.

## References

Barrueco, J. M., & Coll, I. S. (2003). Open archives initiative. Protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. *El profesional de la información*, *12*(2), 99-106. DOI = 10.1145/1667062.1667064. http://doi.acm.org/10.1145/1667062.1667064

Haslhofer, B. and Klas, W. 2010. A survey of techniques for achieving metadata interoperability. ACM Comput.

Hillmann, D. (2003). Using Dublin Core-Dublin Core Qualifiers.

Sicilia, M. A, García E. (2003). On the  Concepts of Usability and Reusability of Learning Objects, International Review of Open and Distance Learning, Octubre, 2003. Retrieved  16th September  2013 at http://www.irrodl.org/content/v4.2/sicilia-garcia.html.

Smith, K., Mork, P., Seligman, L., Leveille, P., Yost, B., Li, M., & Wolf, C. (2011, August). Unity: Speeding the creation of community vocabularies for information integration and reuse. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on* (pp. 129-135). IEEE.
Surv. 42, 2, Article 7 (February 2010), 37 pages.

Tsumoto, S., Hirano, S., & Tsumoto, Y. (2011, August). Information reuse in hospital information systems: A data mining approach. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on* (pp. 172-176). IEEE.

Van de Sompel, H., Young, J. A., & Hickey, T. B. (2003). Using the OAI-PMH... differently. *D-lib Magazine*, *9*(7/8), 1082-9873.

Weibel, S. (1997). The Dublin Core: a simple content description model for electronic resources. *Bulletin of the American Society for Information Science and Technology*, *24*(1), 9-11.