Structured Metadata for Direct Resource Location: A Case Study

Gauri Salokhe, James Weinheimer, Maria Grazia Bovo, Moira Agrimi, Johannes Keizer, Stephen Katz
Food and Agriculture Organization of the United Nations, Italy
{Gauri.Salokhe,Jim.Weinheimer,MariaGrazia.Bovo,Moira.Agrimi,Johannes.Keizer,Stephen.Katz}
@fao.org

Abstract

This paper proposes that for scientific and technical information resources, a well-structured and high-quality metadata record contains enough information to find that resource on the Internet, and as a consequence, no additional human labour is needed to create or maintain any links. Research was performed by creating a control group of records from the Online Catalogue of the Food and Agriculture Organization of the United Nations and searching them in various ways in Google and Metacrawler. Based on results, this method was revised and used on the larger AGRIS database. Results showed not only that the method is successful; it is also highly useful for searching citations. A user interface is suggested, and changes to current cataloguing rules are discussed.

Keywords: *Metadata, Information retrieval, AGRIS, Semantic standards.*

1. Introduction

Although bibliographic databases are rich with structured metadata, and many of the resources they describe may be online, the vast majority of records do not contain Uniform Resource Identifiers (URIs) [1]. The underlying problem with URIs is that they need to be manually assigned to a resource and will not guarantee that the resource can be found at that location over time. Obviously, adding URIs to each of these electronic resources is a tremendous undertaking; therefore, it seems prudent to explore alternative mechanisms to provide access to these resources. Using hard-coded links, such as Uniform Resource Locators (URLs), Document-Object Identifiers (DOIs) or Persistent Uniform Resource Locators (PURLs) to provide the links is equally labour intensive [1,2,3]. In this paper, we propose that structured metadata provides enough information to find a resource on the Web, even if the information in the electronic resource is completely unstructured.

Our study focused on the AGRIS database[1] which has a large quantity of metadata records, dating back several decades, plus a preliminary study on the FAO Online Catalogue, which follows the same basic practices as AGRIS[4]. We felt that if a semi-automatic method could be found to create links from the AGRIS metadata records to the full text resources, it would be a tremendous savings of labour and simultaneously add value to the AGRIS database.

FAOBIB was chosen because it could be limited to resources that are available online. Thus, the reliability of the method depends on the ability to retrieve (or not to retrieve) the resource which we knew to be online. For the purposes of this study, we decided to limit ourselves to finding resources only in English, French, or Spanish[2]. Searching resources in other languages lies beyond the scope of this research, but if diacritics are normalized in Google, results should be similar for other roman alphabets. Non-roman alphabets were not explored at all.

If this method succeeds, it would turn out that metadata records already have sufficient information to find a resource without any extra human-intervention. This new method of connecting the bibliographic record to the resource also has important implications for searching citations, which allows for genuine research on the Web.

In this scenario, metadata becomes essential. The first task of a user is to find a metadata record of a relevant resource, and therefore, consistent, complete, and well-structured metadata is crucial. Our research deals only with a new method to create and maintain access from a record to the full-text.

Several commercial search engines such as LinkFinderPlus[3] and SilverLinker[4] provide the capability of linking metadata records to their full-text resource. However, the twofold cost of paying the search engines and subscriptions to databases and electronic journals may not always be within the budgets. Our method suggests that we can find a full-text resource by generating a string from

[1] AGRIS is the international information system for the agricultural sciences and technology. It was created by the Food and Agriculture Organization of the United Nations (FAO) in 1974, to facilitate information exchange and to bring together world literature dealing with all aspects of agriculture. To date, 240 national, international and intergovernmental centres participate from all over the world. Currently, the AGRIS database contains 3,000,000 bibliographic records. Web Site: http://www.fao.org/agris/.

[2] English, French and Spanish are three of the five official languages of FAO, along with Arabic and Chinese.

[3] LinkFinder*Plus* http://www.endinfosys.com/prods/linkfinderplus.htm.

[4] SilverLinker http://www.ovid.com/site/products/tools/silverplatter/silverlinker.jsp.

structured metadata to query an unstructured full-text database, such as Google.

2. Methodology

This paper assumes that the unique identifier in a metadata record is the title, since even standard numbers e.g. ISBN, ISSN, are not in every record. With scientific literature, titles are normally unique and provide adequate information for locating a resource.

The method proposed was to take the exact title of a resource, as found in the record, and search it in a general search engine, such as Google. Far from implying that a title search is all that is needed to find a resource, the results in this paper underscore the need for high-quality metadata. The hypothesis is that once a metadata record is found, the title will have sufficient information to find the resource anywhere on the Web. There are many caveats here which are discussed below.

2.1. Google

Google is a popular search tool providing a user-friendly online service to millions of users [5]. It is a large-scale search engine which makes extensive use of hypertext and is claimed to crawl and index the web efficiently providing users with highly pertinent results. During the study, it was observed that Google has a number of limitations, such as, a search limit of 10 words, after which the search terms are ignored [6]. It limits the depth it indexes into a directory structure and therefore some resources that may be available online are not found in Google [7]. The placement of cookies on individual machines for page ranking also affects search results [8].

2.2. Metacrawler

Metacrawler is one of the most popular and widely used "meta" search engines [9]. Unlike Google, it does not maintain its own database of information about sites on the Internet. As an alternative, it searches other search engines, such as, About, Ask Jeeves, FAST, FindWhat, Google, Inktomi, Overture, Teoma; and presents a normalized and uniform set of results, providing searchers with the capability to search multiple search engines simultaneously.

Some of the search engines that are being searched by Metacrawler do not implement exact searches, and therefore, even if a search were done with quotes, the results did not always contain the exact phrase. It is also not able to provide the advance search feature of Google, such as, find results with all the words **and** find results with exact phrase.

2.3. Cataloguing Rules and their Impact on Information Retrieval

For our test, the specifics of cataloguing rules—normally a highly esoteric affair—take on critical importance. These rules were developed for the AGRIS network and are not based on other standards, such as *International Standard Bibliographic Description* (ISBD) [10]. Therefore, the structure and information within an AGRIS record can be different from other bibliographic records. Still, there are areas for Titles, Authors, Publication information, and so on.

AGRIS cataloguing rules, also used in FAOBIB, allow for a great deal of latitude. For example, the practice of some libraries in the AGRIS network is to treat the title of the proceedings of a conference simply as ***Proceedings***.[5] This is obviously insufficient information to find a resource. Such titles were excluded from this study.[6]

Another cataloguing rule that could have an impact on information retrieval is that any typographical errors in the title are automatically corrected.[7] There is also the practice of *title enrichment*. This occurs when cataloguers supplement the title, which ***"correct the deficiencies and will reflect the content of the document."***[8] Either of these practices could lead to difficulties finding a resource using the exact title.

English title:	Development of a research programme in irrigation and drainage in Pakistan
Mon.sec.title:	Proceedings of a roundtable meeting, Lahore, Pakistan, 10-11 November 2000
Serial:	IPTRID Programme Formulation Report (FAO/UNDP/Word Bank/ICID/IWMI). 1020-8348, no. 9
Corp.authors:	FAO, Rome (Italy). Land and Water Development Div.
Division:	AGL
Publ.place:	Rome (Italy)
Publisher:	FAO
Publ.date:	Apr 2002
Collation:	164 p.
Languages:	English
Notes:	Summary (En)
IC/IY(2):	XF02
Categories:	F06-Irrigation P11-Drainage
AGROVOC main descr. :	IRRIGATION; DRAINAGE; RURAL AREAS; WATER MANAGEMENT; WATER RESOURCES; WATERLOGGING; SALINITY
AGROVOC geogr. descr.:	PAKISTAN
Publ.type:	D
Job No:	Y3690
Call No:	S238 179

[5] This is also the practice in other cataloguing rules.
[6] A more complex method may be developed later to search conference names in conjunction with a title.
[7] From AGRIS Guidelines (1998), Field 200 Rule 7. In other rules, the practice is to transcribe exactly what is printed, and add an additional title for the corrected title.
[8] From AGRIS Guidelines (1998), Field 200 Rule 8.

Holding library:	LIB
Full text:	English
Acc.No:	408269
Database:	FAOBIB

Figure 1. Sample Record from FAOBIB

The sample record (Figure 1) provides an example of cataloguing from FAOBIB. In this record, there are three titles: *English Title, Title of the Monograph*, and a *Serial Title*. The most unique title here is the English title and therefore was chosen for the study.

2.4. Experiment conditions

We limited the records to items available online in English, French and Spanish, and from these we generated 100 random accession numbers and searched those records. The title of each record was searched in Google and Metacrawler, both as *exact phrase* and as *free-text*. The exact phrase search was carried out with initial and final quotes (“ ”) around the search string while the free-text search involved a search query without the quotes. Therefore, four different types of searches made were: Google Exact, Google Free, Metacrawler Exact, and Metacrawler Free.

Search results were only recorded when they appeared within the top 10 results or when there was a link directly to the resource leading from the first search result. According to this methodology, we could expect that every resource should be found in Google and/or Metacrawler, while the exact phrase search should allow the resource to come up higher in the rankings.

2.5. Analysis of results from FAOBIB

The following graph shows the percentage of resources found for each type of search.

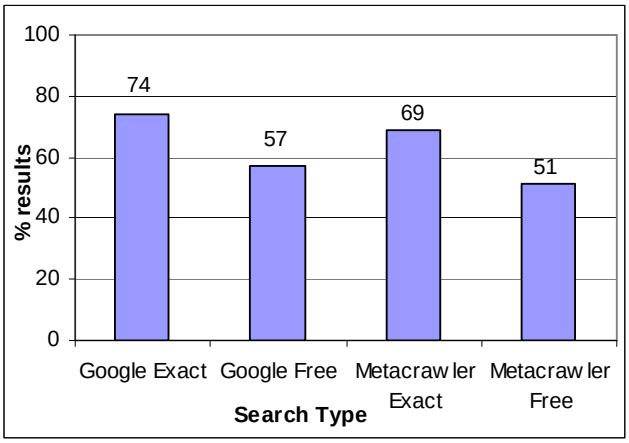

Figure 2. Search results by type of search

Although 100% of these resources are digitally available, we only found 74% with exact search on Google. The lower success rate for free searches, both on Google and Metacrawler, can be explained by the results not appearing in the top 10 limit. Metacrawler did substantially worse than Google.

The following graph shows the percentage of results yielding the exact resource for each of the three languages (English, Spanish, and French) and for each of the four different searches performed.

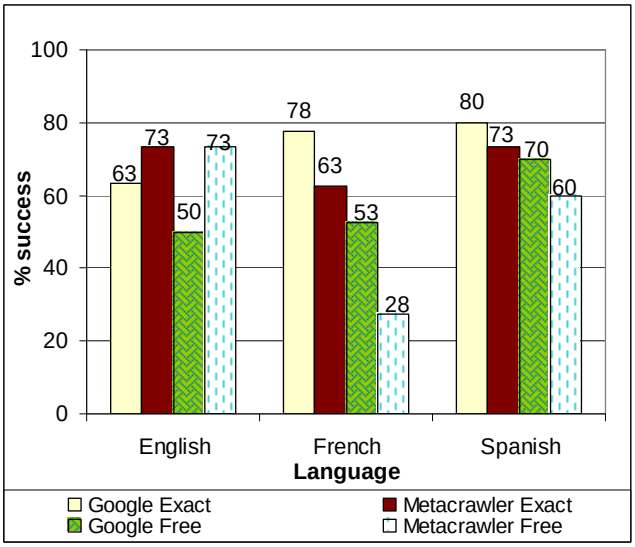

Figure 3. Search results by language of the resource

Figure 3 shows the distribution of resources found by language. Again, Google performs better than Metacrawler. It was interesting that on Google, French and Spanish resources were retrieved at a significantly higher rate than English resources.

There were two cases where the results were beyond our limit of 10. We refined the search adding the name of the first author. This resulted in two hits, and in one case, the total number of results retrieved changed from 797 to 22, with the first result being the actual resource.

2.6. Conclusions from the control experiment

Why was our success rate 74%? There could be several reasons for this. Many of the resources searched had deep directory structures and some of these resources may not have been indexed because Google has a limit to the depth it harvests in directory structures. Also, diacritics and corrections to the title may play a part. Some resources could be too new to yet be in Google. The final conclusion was that our method always found the resource when it was in Google, but Google itself is incomplete. The conclusions from the preliminary test can be seen to justify the chosen method.

In the process, we discovered an excellent possibility of finding citations, which is a primary research tool. A full 80% of the exact phrases searched in Google resulted in finding at least one citation. Metacrawler exact, on the other hand, provided 93% related resources but we discovered

that the exact phrase search was not adequately reliable to be considered a citation [See section 2.2 on Metacrawler].

Taking into account the results presented in Graph 2, we saw no added benefits of searching in Metacrawler, especially since most of the search results from Metacrawler came from Google. Additionally, there was a major discrepancy between exact phrase searching and free-text searching in Google (74% vs. 57%, respectively). Furthermore, the possibility of searching for citations finally led us to choose only Google exact in the case study of AGRIS database.

3. AGRIS

The AGRIS database contains 3,000,000 bibliographic records. The database itself is split into two sections: the *Current* section (records entered from 1996-) and the *Archive* (1975-1995). The system collects metadata for conventional (journal articles, books) and non-conventional materials (sometimes called "grey literature" e.g. theses, reports, etc.) that are not available through ordinary commercial channels. One of the main reasons for the existence of AGRIS is to encourage the exchange of information among developing countries, whose literature would not be covered by other international systems.

3.1. Method

The method used was essentially the same as that used in the first test, except we decided to use only exact title searches in Google, the free text option being seen as less useful, especially for searching citations. Therefore, exact titles were searched, limited to resources in English, French and Spanish, and no generic titles. Hits were counted as those being in the top 10, or if there was a link to the resource from result number one.

We took a random sample of AGRIS records by generating 500 random accessions numbers. Half of the records were in the Current section, and the other half in the Archive. We had no idea what to expect concerning the total number of resources found on the Web. The popularity of placing resources on the Internet did not really take off until 1996, so we hypothesized that more resources would be found in the Current section. Although the Archive contains older materials, and probably has fewer resources online, it could be a rich source of citations to find related documents.

3.2. AGRIS Record Structure & Practice

AGRIS records (see Figure 4) are created according to the rules published in the AGRIS Cataloguing Rules [11]. The title chosen for research here was “Original Title” since it is the title that appears on the resource. The AGRIS database primarily comprises records of scientific and technical articles. This ensures that titles are normally unique.

Accession Number:	97-001686
Title:	Effect of hot water in the germination of Leucaena leucocephala cv."Cunningham".
Original Title:	Efecto del agua caliente en la germinacion de Leucaena leucocephala cv. "Cunningham".
Publication Year:	1995
Subject Category:	Seed production and processing;
Author:	Gonzalez, Y.;Mendoza, F.
ISSN:	ISSN 0864-0394.
Bibliographic Source:	6 tablas; 8 ref. Pastos y Forrajes (Cuba). (1995). v. 18(1) p. 59-65.
Summary lang:	EN ES
AGROVOC keywords:	
English:	leucaena leucocephala; germination; seed; seed storage; water; temperature;
French:	leucaena leucocephala; germination; semence; stockage des semences; eau; temperature;
Spanish:	leucaena leucocephala; germinacion; semillas; almacenamiento de semillas; agua; temperatura;

Figure 4. Sample AGRIS Record

The results of the random accessions numbers led to the following distribution of languages: of 500 records, 338 were in English, 155 in Spanish, but only 7 in French[9] (see Figure 5).

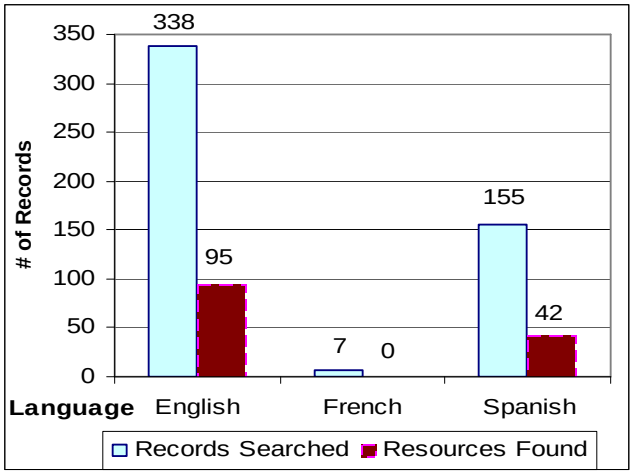

Figure 5. Searches by language

3.3. Analysis of results from AGRIS

The results from the AGRIS test were highly positive. It turned out that the method of searching exact titles in

[9] A follow-up project could concentrate only on French records.

Google resulted in finding 137 of 500 records, or a 27.4% success rate. Just as interesting is the success rate for finding citations: 222 records from 500 found citations to the resource, or 44.4%.

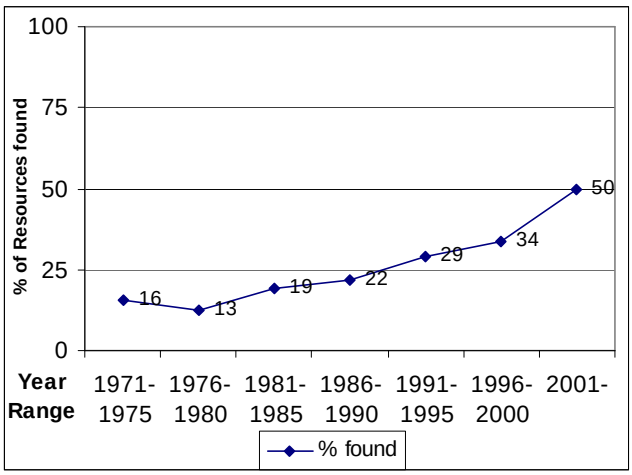

Figure 6. Trend in successful search results by Year

Figure 6 shows the distribution of records searched and resources found by year. The majority of resources found date from 1991. This indicates that there is an effort to put resources on the Internet retrospectively.

If these numbers should hold true for the entire AGRIS database of 3,000,000 records, it would mean that around 840,000 records would have immediate access to the full-text, while 1,400,000 records would provide citations to later documents—all of this without any human intervention. These percentages can only grow with time.

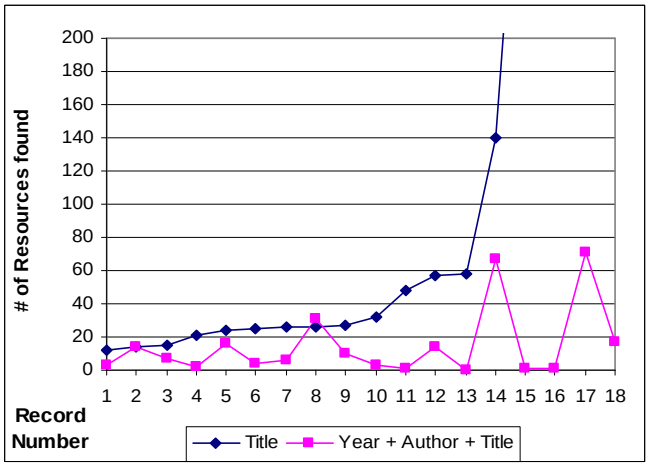

Figure 7. Reduction of noise in search results.

A successful result meant that the resource could be found in the first ten results on Google. Only 18 cases (Figure 7) resulted in an excessive amount of noise, often due to generic titles, e.g. “Dressing and sauces”. When we searched these resources again, this time adding just the year of publication and first author, the noise dropped significantly, in one case going from 285,000 to 1. Some results remained above 50, but in these cases, the articles were important and cited many times.

4. Conclusion

Creating an automatic exact phrase search in Google by taking the original title from a metadata record should lead to approximately 840,000 records in the AGRIS database with access to the full-text. The full-text can exist anywhere in the world. Approximately 1,400,000 records should find useful citations. The Archive, comprised of records created before 1995, turns out to be particularly useful as a source for citations. Utilizing this method allows access to the full-text to be created and updated automatically without any human intervention.

Certain problems remain with incomplete coverage in Google and Metacrawler, and although these are certainly troublesome, the difficulty lies in the search engines and not in the method of searching. Reconsideration of directory structures could solve this problem.

In light of these results, we can visualize a new interface that would allow users to search the AGRIS catalogue just as they do now. When they find a record of interest, they can make an automatic exact title search on Google. If they find the result too large, they will be able to click other parts of the metadata record to add additional information to the search: author, year of publication, series/serial title, etc., whatever they would want to try.

High-quality metadata means adherence to semantic standards (i.e. cataloguing rules) and above all, consistency, but it is important to remember that rules need to change when new circumstances arise. For example, in 1968 people could never have imagined that the titles they entered into their records could someday be searched automatically in a giant computer based in the USA, which would then search the entire world, ultimately finding a digital version of the same resource they catalogued, that had been placed on still another computer in Bangkok, Thailand in 1999. Similarly, they could never have guessed that their rule of automatically correcting typographical errors could make that same search impossible (i.e. a resource has the title “Dre**s**ings and sauces”, while the record has “Dre**ss**ings and sauces”). Therefore, it is clear that the purpose of recording the title has changed. This rule, along with other rules and practices, should be reconsidered in light of new demands and new possibilities.

References

[1] URIs can either be Uniform Resource Locators, Uniform Resource Names, or Uniform Resource Characteristics. For more information, see Uniform Resource Identifier and Uniform Resource Locator. Web site: http://www.w3.org/Addressing/.

[2] Document Object Identifier. Web site: http://www.doi.org.
[3] Persistent Uniform Resource Locator. Web site: http://www.purl.org.
[4] FAO Online Catalogue. Web site: http://www4.fao.org/faobib/index.html.
[5] Google Search Engine. Web Site: http://www.google.com/.
[6] The “Google Hack” of using wildcards was not attempted in this study. Web site: Google Hacks. Getting around the 10 word limit. O’Reilly http://hacks.oreilly.com/pub/h/125.
[7] Restructuring a dynamic site ‘statically’ for max Google crawlage. Web site: http://www.webmasterworld.com/forum3/10804.htm.
[8] Why we target Google? Web site: http://www.google-watch.org/bigbro.html.
[9] Metacrawler Search Engine. Web site: http://www.metacrawler.com/.
[10] International Standard Bibliographic Description. Web site: http://www.ifla.org/VI/3/nd1/isbdlist.htm
[11] AGRIS: guidelines for bibliographic description and input sheet preparation. Rome: Food and Agriculture Organization of the United Nations, Jan. 1998. Available from: ftp://ext-ftp.fao.org/GI/agris/pdf/guidelns/main.pdf.