

INCISO: Automatic Elaboration of a Citation Index in Social Science Spanish Journals

José M. BARRUECO (*), Julia OSCA-LLUCH (**), Thomas KRICHEL (***),
Pedro BLESA (****), Elena VELASCO (**), Leonardo SALOM (**)

Jose.Barrueco@uv.es, m.julia.osca@uv.es, krichel@openlib.org, pbleesa@dsic.upv.es,
elenavelascoarrovo@yahoo.es, leosamu@eui.upv.es

* Biblioteca de Ciencias Sociales, Universidad de Valencia, 46022 Valencia

** Instituto de Historia de la Ciencia y Documentación López Piñero (Universidad de Valencia-CSIC) 46010 Valencia

*** Palmer School, 720 Northern Boulevard, Brookville 11548-1300, USA

**** Dept. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 46022 Valencia

Keywords:

Bibliographic references, Bibliometrics, Citation indexes, Digital libraries, Impact factor, Evaluation of the scientific production

Références bibliographiques, bibliométrie, index de citation, bibliothèques numériques, facteur d'impact, évaluation de la production scientifique

Abstract

Citation indexes are key tools in the science communication system for two reasons. Firstly, they are an excellent information source for searching the scientific literature since they enable navigation through links between documents represented by bibliographic references. Secondly, they allow the evaluation of the scientific production. Citations count is a usual procedure to evaluate the quality of a research paper. In Spain, this evaluation can only be carried out using tools elaborated by the ISI which have a limited coverage of journals published outside Anglo – Saxon countries. In this way, the evaluation of the Spanish scientific production is limited to works published in international journals. There is no tool for the evaluation of research (mainly in Social Sciences and Humanities) published in local journals. With the INCISO research project we will investigate the possibility of create a citation index by automatic means. The deliverable of the project will be software to automatically create citation indexes and a sample citation index for social sciences.

This research is supported by grant HUM2004-05532 from the Spanish Science and Education Ministry.

1. Introduction

Scientific journals are the principal means used by the scientific community to communicate research results. The measurement of the impact of scientific journals has turned out to be a key tool to evaluate the spread and visibility, the significance and importance, and the quality of the research activity. In order to calculate the journals impact factor for a concrete discipline, it is

necessary to build bibliographic databases that must include all the published works on the most important journals within such area. Furthermore, they should contain information about the references cited by each paper in such way that links between citing and cited papers could be traced out. Finally, the system must be able to count the citations that these works have received. Such databases are called citation indexes. At the moment the Institute for Scientific Information (ISI) publishes three indexes covering all disciplines (Science Citation Index, Social Science Citation Index and Arts and Humanities Citation Index). They are the source data for research evaluation in universities world wide.

The high cost and extraordinary technical complexity that involves the creation of citation indexes have inhibited, until now, the development of new databases that could be used as a complement to the ISI products. For non-English speaking countries such a complement is necessary since ISI only deals with international journals, the vast majority of which use English.

In 1983, Garfield alerted about the fact that the population that use to evaluate the impact was mainly Anglo-Saxon. Therefore, any evaluation associated to it has only sense inside this community. Different authors have analyzed the data of the SCI, and have compared it with the scientific production of non Anglo-Saxon countries, and have found that there is a clear discrimination with respect to these countries. This problem can be observed in basic sciences and in technology. But it is even worst in the case of social sciences and humanities, because in these cases researchers use often national or regional journals because these are more connected with the local scope of their research. In this way, research published in local journals is out of the ISI coverage and can not be evaluated.

This project is not the first initiative to develop citation indexes in Spain. Since the 90's several other attempts have been carried out. These were the "Citation Index and bibliometric indexes from Spanish journals of internal medicine and its specialities" (Terrada et al., 1991), the "Documentation citation index in Spanish" (Moya et al., 1998), the "Citation index of Spanish journals of humanities" (Sanz et al., 1998), the "Citation index of Spanish journals of psychology" (Tortosa et al., 2002), the "Citation index of business economics" (Hernández et al., 2003), and more recently the "Index of social sciences" (Jiménez-Contreras et al., 2004). All of them are focused on a specific scientific area, with a delimited time frame and also in the specific geographical framework of Spain.

Most of the projects we cited, unfortunately, disappeared once the research funding was over. All of them share the same characteristics:

- They are based in the work of humans to register manually references and citations
- They are focused on a concrete discipline and
- They use a reduced sample of journals (4-5 in some cases) and include as little information as possible in order to reduce the work load for data typists.

Our conclusion is that the resources required building general citation indexes by traditional means are too expensive to be carried out at national level. In the past only the ISI had the resources to build indexes of printed journals. Nevertheless, with the generalization of the Internet as a new communication channel, with electronic journals that are proliferating both at national and international level and with the possibility of creating indexes by automatic means, new avenues become available. If articles are available in digital formats, there is a possibility for a computer system of extracting the references automatically. With such a system the costs would be dramatically reduced and new indexes covering new document types (e.g. grey literature) could appear.

Trying to further develop this idea and based in the work of the authors described in the next section, we decided to investigate the possibility of developing a computer system which would

be able to automatically create citation indexes for Spanish publications. Our proposal got funding from the Spanish Ministry of Science and Technology with a research grant for three years starting in July 2005. We named the project INCISO (Indice de Ciencias Sociales). INCISO tries to reduce the costs of the process by replacing the human with a computer system that could automatically build an index of electronic journals. It has two main objectives:

- 1) To design a computer system for the elaboration of a citation index in an automated way. The system will have an application to multiple disciplines. Nevertheless, it will be tested with a selection of Spanish journals in Social Sciences.
- 2) To elaborate and disseminate a citation index for Social Sciences based on a selection of Spanish journals. This index will be available for all the scientific community and it will be freely accessible at the project web site at <http://inciso.openlib.org/>.

The remaining of this paper is organized as follows. Section two describes some other research projects at international level which are working in automatic extraction and reference linking to build citation indexes. In section three we analyze the methodology and work plan of our project. INCISO architecture is discussed in section four. Section five describes the status of the project and concludes the paper.

2. Related work

The generalization of electronic formats for publishing and distributing scholarly papers enabled new developments in information retrieval like for example full text searches, reference linking or autonomous citation indexes. Our project is in the last area. Roth (2005) describes several other related research projects that are currently working in developing citation indexes that could be potential competitors of Science Citation Index (SCI). Within Roth's listing we differentiate at least two groups of projects: commercial and academic projects.

Commercial projects usually are carried out by publishing companies to broaden and deepen their bibliographic services. From the technical point of view they use information about documents and references already available in the publisher databases. Such information has been created as part of the editorial process of the documents, usually in the form of tagged SGML or XML documents. The high quality of the data makes possible the development of good added value services. The technical challenges of these projects are the linking of references with full texts across different platforms and the management of access rights to the documents. Examples of such projects are:

- Chemical Abstracts offers cited references searching. Each record is linked to other records, beginning in 1997, that cite it correctly through two features: "Get related" and "Get citing references". The last one allows users to know the number of times an article has been cited. See <http://www.cas.org/casdb.html>
- Scopus is generally considered as a potential competitor to the SCI since it delivers search results that include abstracts, cited references and links to citing references (Roth 2005). Scopus is an initiative of the giant of STM publishers, Elsevier. It has recently added 13 million patent records. See <http://www.scopus.com>
- CrossRef. is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full-text content, but rather effects linkages through Digital Object Identifiers (DOI), that are tagged to article metadata supplied by the participating publishers. The end result is a scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article. It started to work on 2000 when the world's leading scholarly publishers joined to form the Publishers International Linking Association, Inc. (PILA), which operates CrossRef. See <http://www.crossref.org>

Academic projects work with all type of documents available on the Internet. They need to analyze the documents full text in order to extract the references and citations that will be linked from to the documents they represent if available in electronic format. In this case the data is extracted automatically by computer programmes. Its quality varies but in general it is not as good as the commercial projects. The improvement of such technical processes in order to extract better metadata is the key challenge of these projects.

- Citebase allows searching in previously analyzed documents, and obtaining results ordered by impact factor. It was developed into the Open Citation Project, supported by Join Information Systems Committee of U.K. and National Science Foundation of U.S.A.
- CiteSeer is both a software system to extract citations and its implementation to computer science, producing a database containing more than 200.000 documents indexed, with more than two millions references. It has been developed in research laboratories of NEC by Steve Lawrence, Kurt Bollacker and C. Lee Giles, see <http://citeseer.ist.psu.edu>
- CitEc is a citation index for Economics based on electronic documents available in the RePEc digital library. It uses a modified version of the CiteSeer software to reference linking documents which are available in open access (mainly working papers). For each record in RePEc it provides the features: ‘cited by’ when the document has been cited by other papers also available in RePEc and ‘get references’ when the references of the citing paper have been successfully linked to the cited documents. See <http://netec.ier.hit-u.ac.jp/CitEc>
- Google Scholar is a scholarly literature database that includes peer reviewed papers, theses, books, preprints, etc. from academic publishers, professional societies and eprint repositories. It automatically analyses and extracts citations and presents them as separate results, even if the documents they refer to are not online. See <http://scholar.google.com>

3. Methodology and work plan

The authors of this paper have extensive experience in the development of autonomous citation indexes since they have developed the CitEc service described in the previous section. Launching this new project tries to export the experience acquired in a concrete discipline to publications in a language different of English, coming from multiple disciplines but with the common indicator of being published in the same country. Most of the software developed for CitEc is going to be used and tested in this new environment.

The methodology we are going to use in order to extract and link the information about references can be described in the following seven steps:

1. We need to select data sources. The system will be tested with a sample of Spanish journals in social sciences. In a first stage this sample is reduced to ten journals representing all disciplines. The selection was carried out taking into account the following criteria: journals should have an electronic version with at least four issues published and a peer review system to assure the quality of the contents. Since the index is going to be created automatically, the requirement of electronic versions of the journals is crucial. The number of electronic journals in Spain is still small. Nevertheless is growing fast as shown in ‘Directory of Spanish Electronic Journals of Social Sciences and Humanities’ (available at: <http://citas.uv.es/DifusionRevistas/Revistaselectronicas/index.html>). There are new journals born only in digital format and some others that are migrating to the electronic

environment but maintaining a printed version. The selection based in the availability or not of electronic versions implies that important journals will be kept out of the sample because they continue being published only in paper. In this way, we are aware that the selection does not include the best Spanish journals and the results should be taken carefully and not used for research evaluation purposes. This limitation is going to disappear in the future as more journals go digital.

2. We need to obtain bibliographic information about the articles published on the selected journals. In the future is desirable to work with information suppliers (publishers) in order to define automatic means to feed information into the system. That means working in procedures which allows INCISO to be aware of new papers published. For this purpose we will use new technologies in the area of digital libraries, such as the protocol OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting), see <http://www.openarchives.org>.
3. The bibliographic information for each article with the electronic address pointing to the documents full text is stored in a MySQL database. These documents will be considered the citing documents. Another table in the database is filled with metadata about documents published in Spain in the social sciences areas on the last ten years. These are the potentially cited documents. This metadata is defined as authoritative since it comes from quality sources. Only citations to such documents will be taken into account and considered as true citations. All other citations will be discarded.
4. For each citing document, the file containing the document full text is downloaded. At the moment INCISO only deals with files in PDF format. The file is converted to ASCII so that the text could be easily extracted and manipulated.
5. Once the file has been successfully converted starts the parsing of the whole text in order to identify and delimit the references section. If this step concludes correctly, it is necessary to further identify each one of the references cited and split it in the different elements that made it up, e.g. author, title, publication, etc. This is the most important part of the system, as the effectiveness of the process depends mainly on the quality and consistency of these results. One of the main problems is the fact that references are different for each discipline. The approach followed by most projects described in section two is to extract all reference elements, as correctly as possible. Thus, they tried exhaustive parsing of the reference. We believe such parsing is complicated and resource consuming since the quality of the source data varies considerably. Our approach is different. The system will only identify basic elements of the reference and then it will try to locate the document referenced in the database of authoritative metadata. If it is able to locate the document, then the reference is completed with the correct and exact metadata.
6. All the data extracted in the previous steps is stored in a database of references. This database will be used for bibliometric studies of the results.
7. The project will offer two types of results. On one hand, the citations index that will be useful for evaluation of the research carried out in Spain in social Sciences, and on the other hand, a set of technical documents about the system that will be of mayor interest for the researchers community on the digital libraries area. All results will be freely published on the web.

INCISO will develop a computer system to carry out the process described previously in an automated way. The design of the system will be based in the following basic characteristics:

- Multi-discipline. Initially the system will be applied to social sciences journals. Nevertheless it will be build on the basis of a modular architecture that will allow easy adaptation of new functions to the nucleus, in order to give solutions to new requirements of different disciplines.
- Based on open source software. The system will be completely written in Perl. The additional software required will be open-source programmes, i.e. software using GNU or similar licences. The system will operate on a Debian GNU/Linux machine based at the Universidad Politécnica de Valencia (Spain), using MySQL as the database management system and Apache as the web server.
- Autonomous and continuous. One of the main requirements to take into account in the design of the system is that it should work with the minimum of maintenance as possible. Current systems are based in the editorial work of administrative people which make necessary the monetary resources to pay them. If we build a system with a maximum of automatic processes, and we are able to obtain a critical mass of documents it would be possible that some publishers may contribute to it their publications. That will assure a continuous flow of documents, and the system could work by itself with the initiative of publishers.
- Open. Data generated will be accessible for all the academic community and for other projects at international level too. The first expansion of the project could be to journals published in Latin America. Latindex, see <http://www.latindex.org>, is a directory of electronic journals compiled by the CINDOC. It could be used to select quality journals to be included in INCISO.

5. System Architecture

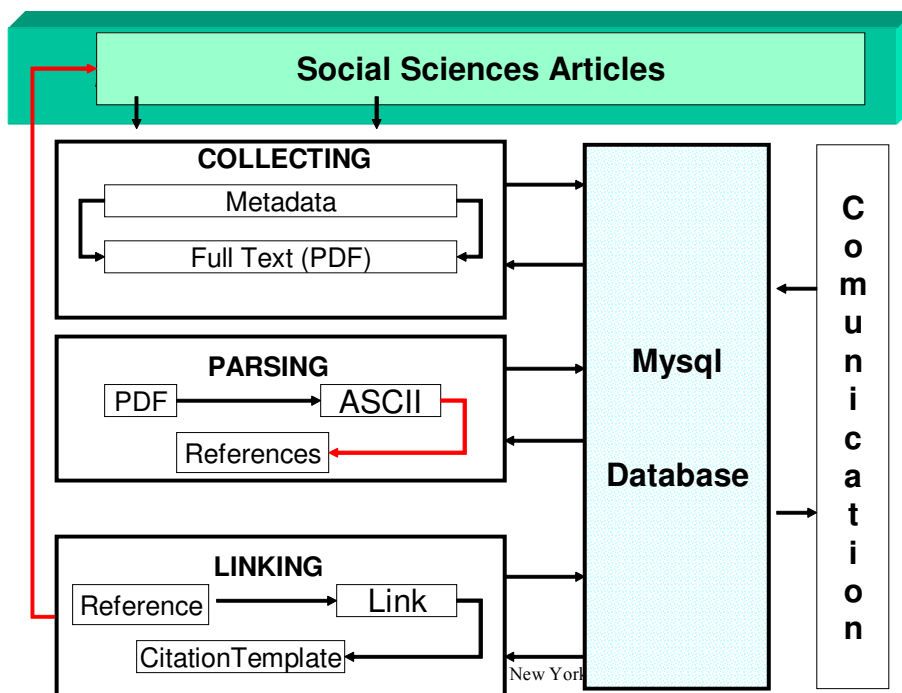


Figure 1: INCISO architecture

As is shown in Figure1, the INCISO architecture is based in two main elements. Firstly, the environment on which we work is made of articles published in Spanish journals on Social Sciences. We have built a databank of authoritative metadata describing each one of the articles. This metadata is stored in a bibliographic database. The precise details of this base are beyond the scope of this paper. Secondly, we have a series of three software modules, one for each step in the reference linking process (Barrueco, 2005):

1. Collecting metadata and documents' full text.
2. Parsing of documents in order to find the references section, to identify each reference and to extract their elements (authors, title, etc.).
3. Linking of references with the document they represent if available on INCISO.

It is important to note that each module is based on the output of the previous one. In this way, the successful processing of each document implies to successfully pass the sequence of three levels. Each document has a status associated with it. The status indicates in which moment of the process is. The initial status for each paper is "nofulltext" and the last one when everything goes successfully: "linked".

1. **Collecting.** Collecting involves three different steps: (1) to collect the documents' metadata, (2) to download the documents' full text and (3) to convert them to a format ready to be parsed by a computer system. Metadata for the citing documents is completed with the URLs of the articles' full text. In some cases the URLs provided may be wrong or the web server may be down when the system tries to access the documents. In such cases articles are marked with a special status name, and the process stops until the editorial staff checks and corrects the problem found. Once the full-text file is saved in our hard disk, we start the conversion process. First, we check if the full text file is compressed. If that is the case, a decompression algorithm is used. Second, we check the file format. Only PDF documents are accepted at the moment. Fortunately PDF is a quite popular format for publishing scientific papers on the Internet. The last step is to convert the document from PDF to ASCII. For this purpose, we use the software pdftotext developed as part of the Xpdf viewer. Not all PDF files can be correctly converted to text with enough quality to allow text extraction. Mainly the quality of the PDF files depends on the software used to create the files and the correct use of font codification.
2. **Parsing** is the most complicated stage. Authors usually construct references in a variety of formats, even within the same paper. In addition disciplines vary with respect to the traditions in the way citations are marked in the documents. Due to the importance of the parsing process we decided to start with software already tested rather than develop new software from scratch. Our choice was the software developed for the CitEc project, which has been described in papers like Lawrence (1999). CitEc software is able to identify the part of the document containing the list of references. Then it can split the list into different references. Finally it parses each reference to find the elements. At the moment it only identifies the publication year, the title and the authors. However, these four elements are enough for our purposes. The quality of the bibliographic references provided in the source papers is variable. For instance, it is usual to find different name forms for the same author, different name forms for the same journal, etc, within the same paper. We use the authoritative metadata to complete and improve the references' quality with authoritative data provided by the publishing institutions.
3. **Linking.** Once we have parsed the documents, the next stage is to look if some of the references successfully extracted go to documents available in the INCISO database. In such cases, some type of link between both documents should be established. We are doing that by comparing each reference successfully parsed, with the authoritative

metadata stored in the INCISO bibliographic database. At the moment we consider that a reference represents an INCISO document when:

- The parsed reference title and the title in our metadata collection are close enough.
- The publication year of both items is the same.
- At least one of the papers authors' matches the authors of the metadata record.

In this process we take each reference, extract the parsed title and convert it to a normalised version called key title. Here all multiple spaces and articles are removed; and all upper case letters are converted to lower case. Then we select from our bibliographic database all documents that contain in their title all the words of the reference key title. All selected papers are candidates of being the cited document. In a second step we compute the Levenshtein distance of each candidate's key title with the reference key title. If this distance is greater than 8% of the reference key title length, the candidate document is rejected. Finally, we check if the publication year of the candidate papers and the reference is the same. If this is the case we assume that the reference is to the document we have. Authors are only compared when the title length is small and it does not discriminate enough. Information about citations is stored in a table of the mySQL database. That database will be used to develop bibliometric indicators.

6. Conclusions

In this paper we have described a methodology to automatically develop a citation index. With the implementation of this methodology the INCISO project will try to reduce the high costs of developing citation indexes by traditional means. If successful, it will open a way for non-English speaking countries to develop their own indexes that could be used as a complement to ISI's for research evaluation.

At the moment we are in the very beginning of the software development. The first results are expected to arrive in 2006. Then a period of evaluation will start in order to determine if the results are good enough to allow both information retrieval and extraction of bibliometric indicators.

There are other projects at international level working in the same area. The innovation of INCISO is the work with a database of authoritative metadata that can perform of the normalization of references extracted from documents.

References

Barrueco, José Manuel, and Thomas Krichel (2005) "Building an autonomous citation index for grey Literature: RePEc, the economics working papers case" *The Grey Journal, An International Journal on Grey Literature*, vol. 1, no. 2, pp. 91–97

Lawrence, Steve, Kurt Bollacker, and C. Lee. Giles (1999) "Indexing and retrieval of scientific literature", proceedings of eighth International Conference on Information and Knowledge Management, CIKM99, pp. 139–146.

Roth, Dana L. (2005) "The emergence of competitors to the Science Citation Index and the Web of Science", *Current Science*, 2005, vol. 89, no. 9. pp. 1531–1536.