Implementing the additional knowledge in the Croatian Scientific Bibliography

Tomislav Jagušt[*], Jadranka Stojanovski[**] and Mirta Baranović[*]
[*] University of Zagreb, Faculty of EE and computing / Department of Applied Computing, Zagreb, Croatia
[**] University of Zadar, Zadar / Ruđer Bošković Institute, Zagreb, Croatia
tomislav.jagust@fer.hr, jadranka.stojanovski@irb.hr, mirta.baranovic@fer.hr

Abstract - The Croatian Scientific Bibliography - CROSBI is a bibliography web site that stores information about scientific papers published in Croatia in the last 15 years. At the present, CROSBI consists of over 400,000 records and is daily accessed and updated by a large number of researchers. As the appearance and purpose of database (and hence its structure) changed over the years, some inconsistencies emerged, and existing data could not be used to its fullest potential, e.g. linking papers from the same journal or conference, or analysis of scientific activity in relation to the place of article publication. This paper explores the possibilities of improving and enriching the CROSBI database, using different techniques, from data cleaning and sanity checking, to creating links between CROSBI and similar online bibliographic databases. In order to improve data quality, book records from CROSBI were linked to the corresponding Google Books records, and using Semantic Web technologies, a number of links between CROSBI and online bibliographic databases was created. The resulting data set can be used to further improve the quality of CROSBI by creating a recommender system, or to improve and open up new possibilities in the analysis of existing data.

I. Introduction

The need to collect and organize data about scientific research in Croatia arose early after gaining independence as a country in 1992, when it was realized that data about scientific output are scattered across different index publication, subscription databases and local bibliographies. Small but regular support from the former Croatian Ministry of Science and Technology (today Ministry of Science, Education and Sports) enabled the start of the first phase of the Croatian Scientific Bibliography – CROSBI[1] development, and first release in 1997 [1]. CROSBI is designed, created and maintained at Rudjer Boskovic Institute Library in Zagreb. In the beginning, the primary goal of CROSBI was to collect the data on scientific output of the current research projects financed by the Ministry of Science and Technology and to make them publicly available [2]. The input was provided using a simple web interface, and monitoring and curation was organized by librarians. CROSBI was designed as a single entry point for various publications as books, book chapters, journal articles, conference papers, theses, reports, manuscripts, patents etc., but was also open for papers not yet published like submitted journal articles (“in press”), unpublished manuscripts, and other types of intellectual products related to the academic and research community. CROSBI served also as an Open Access (OA) repository offering free access to deposited publications long before OA movement became turning point in the scholarly publishing area. The data stored in CROSBI was used mostly for an ad-hoc insight into a comprehensive bibliography of an institution, research project or individual researcher [3].

Today, CROSBI consists of 410,000 records, about 130,000 of which are bibliographic data about journal articles, and more than 150,000 are about conference papers. Furthermore, CROSBI provides an access to 25,000 records with full-text available [4]. Authors/Researchers themselves enter the data into the database, so there is a high level of up to data.

The CROSBI system is based on open source solutions: Linux, PostgreSQL DBMS (replacing mSQL) and Apache. PostgreSQL is used for database management because it is unified database server with a single storage engine, with reliable and fast performance in complex operations, and the Perl interface used in cgi-scripts on the Web server. Apache web server is used together with SSL encryption mechanism to provide HTML content to end-users.

During the years, some inconsistencies in the database emerged, like data glitches, incomplete or missing data, missing lookup tables, etc. Since incorrect or inconsistent data could complicate database maintenance, prevent the exploitation of data to its fullest potential and lead to false conclusions, we decided to explore the possibilities of improving the quality and enriching CROSBI data, by implementing different approaches and techniques, from ETL-like procedures to the Semantic Web technologies. The goal is to enhance CROSBI from a stand-alone database of bibliographic records to a highly hyperlinked data set that can interact with other information resources on the web. Development of CROSBI will follow the principles of Linked Data, with resolvable URIs for all objects, ensuring structured RDF data. Using rich ontologies and semantic linkages data stored in CROSBI will not only be available on the web, but will really become an integral part of it.

In the future CROSBI is seen as an interoperable tool, exchanging data with others repositories, archives and databases, which will enable much easier and controlled

[1] http://bib.irb.hr/

This work was partially supported by the "Semantic integration of heterogeneous data sources" project, funded by the Croatian Ministry of Science, under grant number 036-0361983-2012.

input process and provide enriched data beyond bibliographic level. The tools and methods applied in this research and presented in this paper will enable such kind of transformation.

II. Current database structure

There are several inconsistencies and shortcomings of the current database structure, which are mainly the result of its revisions and adjustments to different needs during the years.

Some data which is a part of CROSBI is scattered over different databases. Most of bibliographic information about publications is stored in one database, but data about authors and research projects are in separate databases. Although located on the same server and within the same database management system, the fact that they are separated into different databases is an obstacle that prevents data integration and consumption of CROSBI as a whole, e.g. there are no foreign keys, and it is not possible to create advanced web services that expose data for public consumption, like RDF graph, since tools that provide these features cannot work with multiple databases at once.

The database is partly denormalized, e.g. table “Proceedings_article” contains information about the papers in the conference proceedings, together with information about proceedings itself, and also with information about the related conference. Similarly, table “Journal_article” contains information about the papers in journals combined with information about journals.

Many attributes (about 70%) in the database relations are of type *character varying*, and web form fields used to enter or edit data are, or at least were, in some point in the past, simple HTML text input fields. Consequently, database columns which are mandatory and can not be skipped during the entry process, contain a lot of inaccurate data, missing data, incomplete data or typographical errors (e.g. attribute “year” contains values like “-”, “not yet published”, “in press”, “2011/12”, “2012g.”, “20111” etc.). Furthermore, there are same entries written in slightly different versions (all caps, camel case, with or without Croatian diacritics). Fields that were left empty by the author were not stored into the database as null, but rather as empty string, or contain some sort of “missing data” label, presented with “-“ or “n/a”.

Many of these discrepancies were corrected during the years, mostly through improved web interface, data-type or range constraints, and better control of user entries, but a certain amount of data entry errors are still present in CROSBI, mostly among the old entries.

There are columns that contain non atomic values, and are in conflict with the First normal form. Place of book, proceedings publication and conference location are two most prominent cases of these fields, where users often entered more than one value, e.g. “Rijeka - Opatija” or “Zagreb, Hrvatska”.

III. Data transformation and cleaning

To clean and ‘sanity’ check the unstructured free-text data, a set of custom ETL procedures was prepared and executed as described in **Error! Reference source not found.**.

A copy of current CROSBI database was made, and used as a test model to explore the possibilities of improving the database content, enabling advanced search and data analysis options, increase the user experience and ease the interconnection possibilities with other similar databases.

First, the database was normalized, which resulted in creation of several new relations (“Language”, “Proceedings”, “Event”, “Theses type”, etc., and their appropriate junction or linking tables). In total, 22 new relations were created.

The next step was to load data into newly created tables. While some lookup tables were loaded from other sources (e.g. a list of universities and faculties in Croatia was loaded from ISVU[2]), most of tables were filled with data from the old denormalized tables, mainly by selecting distinct values, and then manually cleaning and separating only the relevant records.

Some tables required extra transformation effort, e.g. a relation “Place” was first loaded with distinct values from all relations containing attribute “place”, then the non-atomic values were split, filtering out country names and removing the unnecessary characters (blanks, commas, semicolons). The last step was to unify alternate city names (e.g. “Beč”, “Vien” and “Wienna” with “Vienna”), and repair some frequent typos, like “Zabreg” or “Zabeg” with “Zagreb”.

The data about proceedings and conferences was extracted using different text similarity approaches. Database table “Proceedings_article” was split into several tables, namely “Proceedings_paper”, “Proceedings” and “Event”. The original table contained 146,628 records, and newly created tables have 125,779, 68,834 and 53,227 records, respectively.

TABLE I. shows comparative overview of the number of records initially loaded into newly created tables and a number of records after the different data cleansing techniques were performed.

TABLE I. A number of records in newly created tables before and after data cleansing

Table	*Starting noumber of records:*	*Number of records after cleansing*
Language	223	36
Proceedings	79,897	68,834
Event	86,572	53,227
Place	3,160	1,159

[2] http://www.isvu.hr/

IV. LINKING CROSBI WITH EXISTING ONLINE SOURCES

In order to further improve the quality of data in the database, records from CROSBI were matched against a number of different available datasets. Three different techniques were applied: offline matching (for datasets which were available as a whole database dump or in similar form), writing a custom (SOAP or REST) web service client, or matching using the semantic web technologies and available SPARQL endpoints.

A. OFFLINE MATCHING

Offline matching was conducted with data from Croatian Higher Education Information System (Croatian acronym: ISVU) and GeoNames[3] geographical database. Database dumps of publicly available data were obtained, and loaded into the CROSBI database. Data matching was conducted in several iterations, each iteration decreasing the number of remaining unmatched records. On the other hand, every subsequent iteration was computationally more demanding and its results contained more errors (mostly false positives).

For instance, places from CROSBI were matched against GeoNames database in the following order: at first, matching was restricted only to Croatian towns with more than 15,000 inhabitants, and the whole string was matched. Some frequent places that do not meet those criteria (like Opatija and Cavtat) were added matched manually at the same time. The remaining data was then matched against a list of big European cities, then against a list of all Croatian cities, and against the list of alternate city names in the whole World. These steps connected about 90% of cities, with very good accuracy, from around 60,000 rows in the beginning; only about 5,000 were still unmatched. The last step in a row was fuzzy string matching (matching only part of the word, Trigram[6], Levenshtein and Soundex[7] techniques), but since the unmatched data contained a lot of erroneous data (like “Unknown city” or “Whole Europe”) the number of hits was small, with a lot of false positives.

B. ONLINE MATCHING - CUSTOM WEB SERVICES

1) Google Books

Google Books[4] is an online service that allows its users to search full text of more than 30 million books and magazines. In addition to web search, Google Books exposes a Representational State Transfer (REST) Web Service - Google Books API, which allows integration of third party applications with Google Books database. The custom REST Web Service client was created, and CROSBI book data was matched against Google Books database. About 1,500 records (of 10,000 in CROSBI) were matched. Since only full title was matched, the number of hits could be further improved by implementing partial matching or some other sort of fuzzy search.

2) Google Scholar

One of the biggest online resources of scientific papers is Google Scholar[5], a free web search engine that indexes the full text of scholarly literature across many disciplines and sources. Although it is free for personal use, Google Scholar does not provide an API, and use of automated data crawlers is in conflict with their usage rules. Matching of a small sample of CROSBI data against Scholar search engine yielded some very good results (around 50% of articles were successfully matched). In the future Google will hopefully include Scholar into its API, and CROSBI could be interlinked and enriched with additional data from Scholar’s huge database.

3) Web of Knowledge

Another big online resource of scientific works is Web of Science[6] (WoS), academic citation indexing and search service by Thomson Reuters. There is a SOAP Web Service available to all subscribers, which allows querying their database by different criteria, and retrieving a set of basic information about discovered articles. A part of CROSBI records that had “Current Content_index” and “Science Citation Index_index” flags set, was checked against WOK database, and about 65% of records were successfully matched.

C. ONLINE MATCHING - SEMANTIC WEB TECHNOLOGIES

Online libraries and bibliographic databases like CROSBI can gain much benefit from The Semantic Web technologies, like Ontologies (specifically Web Ontology Language - OWL) and SPARQL Protocol and RDF Query Language (SPARQL)[8]. Since some of the online bibliographic databases, search engines or Web sites have a semantic version, accompanied with a SPARQL endpoint (a REST Web Service that accepts SPARQL queries and returns results in a simple XML form), a custom Semantic tool was made to match CROSBI data against different SPARQL endpoints. In our research, PubMed (Medline), DBLP, ACM, CiteSeer and Jucs.org were queried with a subset of CROSBI data. Taking into account the restrictions of endpoints, response and computation time of text similarity functions in SPARQL, only simple text search was performed on most of the endpoints. The results are shown in TABLE II.

TABLE II. DATA MATCHING AGAINST DIFFERENT SPARQL ENDPOINTS

Endpoint	***Url:***	***Matches***
PubMed	http://pubmed.bio2rdf.org/sparql	5567
DBLP	http://dblp.rkbexplorer.com/sparql/	1513
ACM	http://acm.rkbexplorer.com/sparql/	168
CiteSeer	http://citeseer.rkbexplorer.com/sparql/	51
JUCS	http://jucs.org:8181/d2rq/sparql	5

[3] http://www.geonames.org/

[4] http://books.google.com/

[5] http://scholar.google.hr/

[6] http://www.webofknowledge.com/

V. Results

Different database modifications, transformations, and data matching against data from various online bibliographic and library sources were performed. The CROSBI database was normalized, most of the inconsistencies and deficiencies were removed, and data accessibility and richness was improved. Some of the results are shown in TABLE I. and TABLE II.

Since the new database tables were created, the possibilities to use or analyze data in a different way emerged. Some examples of that would be: analysis and evaluation of Croatian universities by a number and type of papers published in different journals or conferences, analysis of scientific production by region or city, and similar.

VI. Conclusion and future work

Conducted activities improved the content of the CROSBI database and opened up new possibilities to use and analyze its data.

Links with bibliographic data on different online databases stimulates the possibilities of further data discovery and enrichment. Since the used online databases, as well as CROSBI, are being constantly updated with new data, data matching should become a regular periodical activity, so newly added records would also get interconnected and enriched.

The existence of links with other databases allows cross checking the CROSBI data with a validated data sets. Also data enhancement, by adding missing or related information, was made possible.

Although there is still a rather small number of Semantic sources available and the number of matched records is, for now, relatively small, the Semantic web technologies showed some significant advantages over the other used technologies. The Web Service Client that consumes SPARQL endpoint needs minimal adaptation to work with different endpoints, data format is open and well defined, and interconnection with other sources or inclusion of additional knowledge into existing bibliographic database is simple by using the *owl:sameAs* property.

The future work should further explore the possibilities of more advanced text comparison methods, to improve the matching efficiency and even more increase the data quality. Also, a newly discovered and collected data could be utilized in a recommender system implemented inside a CROSBI web interface, which would help and guide users in correcting and improving the data about their papers and books.

Acknowledgment

We would like to express our special thanks to CROSBI developer prof. Ivo Batistić, from the University of Zagreb, Faculty of Science, for creating the database system and investing a lot of hard work in the maintenance and improvements of CROSBI in last 15 years. Special thanks go also to system's librarian Alen Vodopijevec from Ruđer Bošković Institute, for continuous technical support to the CROSBI bibliography and provision of data needed for this research.

This work was partially supported by the "Semantic integration of heterogeneous data sources" project, funded by the Croatian Ministry of Science, under grant number 036-0361983-2012. The authors would like to thank the funding agency for their support.

References

[1] Stojanovski, J., & Slavic, A. (1999). Electronic bibliography - its reliability and its impact on the concept of bibliography in general. In T. Aparac, T. Saracevic, P. Ingwersen, & P. Vakkari (Eds.), Conference on Conceptions of Library and Information Sciences - COLIS (pp. 1–10). Lokve, Croatia: Benja Publishing.

[2] Stojanovski, J. (1999). Bibliography in the network environment: Croatian Scientific Bibliography (CROSBI). In Information Technologies Interfaces ITI 1999.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Stojanovski, J. (2002). Croatian Scientific Bibliography (CROSBI) - four years experience. In A. L. Markham, James W. ; Hyett, David J. ; Duda (Ed.), Managing resources in a sea of change: Proceedings of the 27th Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centres (IAMSLIC) and the 9th Conference of the European Association of Aquatic Scien (pp. 65–75).

[4] Croatian Scientific Bibliography - Total statistical data http://bib.irb.hr/skupni_podaci?lang=EN, Accessed: February 24, 2014.

[5] H. Muller, J. Freytag (2003), Problems, methods and challenges in comprehensive data cleansing, Humboldt-Universitt zu Berlin, Institut fr Informatik, Berlin

[6] PostgreSQL 9.3 Documentation: pg_trgm module, http://www.postgresql.org/docs/9.3/static/pgtrgm.html, Accessed: February 21, 2014.

[7] PostgreSQL 9.3 Documentation : The fuzzystrmatch module, http://www.postgresql.org/docs/9.3/static/fuzzystrmatch.html, Accessed: February 21, 2014.

[8] Svensson, L. G. (2013). Are Current Bibliographic Models Suitable for Integration with the Web?. Information Standards Quarterly, Winter, 25(4), 6-13.