

Links and Hubs of Scientific Information - the Case of the MTMT

András Holl¹, and Gábor Makara²

¹ Konkoly Observatory of the Hungarian Academy of Sciences, Budapest, Hungary;

Library of the Hungarian Academy of Sciences,

`holl@konkoly.hu`

² Institute of Experimental Medicine, Hungarian Academy of Sciences, Budapest, Hungary

`makara@koki.hu`

Abstract. We present the Hungarian National Scientific Bibliography project: the MTMT. We argue that presently available commercial systems cannot be used as a comprehensive national bibliometric tool. The new database was created from existing databases of the Hungarian Academy of Sciences, but expected to be re-engineered in the future. The data curation model includes harvesting, the work of expert bibliographers and author feedback. MTMT will work together with the other services in the web of scientific information, using standard protocols and formats, and act as a hub. It will present the scientific output of Hungary together with the repositories containing the full text, wherever available. The database will be open, but not freely harvestable, and only for non-commercial use.

Keywords: bibliography, scientometry

1 The MTMT

1.1 The Project

The Hungarian National Scientific Bibliography project (MTMT, abbreviated from the Hungarian name) aims to create a comprehensive national scientific bibliography of all researchers working in Hungary, affiliated to Hungarian institutions, funded by Hungarian grants or participating in the scientific life of the country in other ways. It will contain publication meta-data and the meta-data of the citations these works received. MTMT is already working¹ as a consortium of 21 members, universities and other centres of research, including the Hungarian Academy of Sciences (HAS), but is expected to be re-engineered with the help of a government grant.

1.2 Rationale

Effective science management is not possible without scientometry. Large international companies offer databases and services for this purpose, but these i.) do not cover the humanities properly; ii.) do not cover books and non-journal literature comprehensively, iii.) do not attribute well the publications to authors and institutions, and iv.) do contain errors, which are hard to correct. For a national scientific bibliography, author or institutional feed-back is necessary, and such a project must cover the fields that publish in the national language. Communication with the large commercial databases is essential nevertheless.

As most of the research is done in multi-institutional teams, and published in multi-author publications, and as individual researchers have multiple affiliations (both in parallel and sequentially), bibliographic databases are not effective when produced at an institutional level. Maintaining these at researcher level is laborious and error-prone. Our solution for this problem is a nation-wide database, with institutional and individual control possibilities. An important

¹ The formal starting date of the project is July 1, 2009.

goal is to minimise the workload on the researchers, while still allowing them control over their own publication data. So the system will stand on three pillars: automatic data import wherever possible, expert bibliographer/librarian control and validation over the data and personal control opportunity for the researchers.

The MTMT can also serve as an aggregator and search engine for Hungarian research. Even though it will neither store full text publications, nor have the full functionality of a Current Research Information System, it can provide links to the articles and documents, and provide bibliographic data for a national R&D information system.

1.3 Current Status of the Project

The MTMT was created by merging two former databases of the HAS, which had had a narrower (and somewhat overlapping) scope. The software was inherited from one of these databases. Although it can (and will) function unchanged in the near future, the present system was not designed to serve the whole scientific community of the country. The development of new software will commence soon, hopefully.

The present MTMT contains over half a million publication records and about 1.8 million citations at the moment, and these numbers are expected to grow with the migration of the local databases of universities. The number of active users is expected to reach 30 thousand in a couple of years.

The web interface [1] is bilingual (Hungarian and English), but at the moment is only available in Hungarian. The English version will be switched on in a year's time, when the data coverage will be close to uniform for all the participating organizations.

1.4 Similar Projects Around the World

There are several nation-wide repositories, research information systems around the world, though these differ to some extent in their objectives, functionality and scope. The National Academic Research and Collaborations Information System (NARCIS) [2] in The Netherlands, the Croatian Scientific Bibliography (CROSBI) [3] could be mentioned as examples.

2 Links to Other Services

2.1 Links to and from Other Services

The data provided by MTMT will be used by other services, e.g. subsets of bibliographic lists of researchers will be used in the grant application or habilitation process. As the administration of these applications can often (or increasingly, can only) be done electronically, the supporting bibliographic information will be provided by MTMT as a web service. We expect to create Service Level Agreements with the partners using MTMT data.

The library world will expect information from the MTMT - for this we plan to implement OAI-PMH and Z39.50 interfaces, and support Dublin Core and MARC meta-data formats. For the data input side MTMT will need information from various providers, and this makes necessary the use of the API-s of different bibliographic databases, and the use of OAI-PMH for harvesting repositories.

As it can be seen in the examples above, the scientific information web consists of several services or hubs, connected with standard protocols. We will explore some use cases further.

2.2 The Case of the Hungarian Scientific Research Fund and the MTMT

The Hungarian Scientific Research Fund (OTKA) grant application is aided by a web-based tool. Applicants need to supply a list of their recent papers, which will be provided by MTMT automatically. The articles regarded most relevant should be marked by the applicant. The reports required mid-time and after the completion of the project pose a similar exercise, only in this case the relevant papers are selected using the grant identifiers stored in the bibliographic record of the MTMT system.

OTKA has an Open Access mandate, so links to the freely accessible copies of papers will be provided by MTMT too - helping to monitor the compliance to the policy. Project reports are archived and published in the Repository of the Academy's Library (REAL).

2.3 The Case of the Institutional Repositories and the MTMT

Repositories will be harvested by MTMT using OAI-PMH. To facilitate proper author and institutional assignment of papers, and the filtering of duplicate copies, a slightly extended meta-data set might be necessary compared to plain Dublin Core. MTMT compatibility of repositories will be certified by the HUNGarian Open access Repositories group. Some participants in the MTMT project prefer the reverse route: their authors will enter the meta-data of their papers into MTMT, which in turn will provide a way for uploading the full text copy to the institutional repository, using the SWORD protocol.

The MTMT will provide links to the full text version of the publication, including the Open Access version, if available, and will display the Open Access status.

2.4 The HAS DSc Degree and the MTMT

The Hungarian Academy of Sciences (HAS) awards a DSc degree. Applicants need to submit a thesis, and certain scientometric parameters are required. Certified bibliometric data are provided by the MTMT. The submitted theses, together with other documents of the process are made public in the repository of the Library of the HAS. This repository, REAL, is harvested by the MTMT in turn, so the theses will automatically appear in the personal bibliographies (which could be embedded in personal web pages too).

2.5 Data Services for the Hungarian Doctoral Council

The Hungarian Doctoral Council (HDC) - an organization created by the universities offering PhD programs - maintains a website offering various information on the doctoral schools, research topics, and academic staff, including bibliographies of the latter - provided by the MTMT. The HDC information system already taps into the MTMT database - but we plan to transform the existing channels of data transfer to web services.

3 Architecture of MTMT

The planned architecture for the new software of the MTMT will consist of a core server and clients used at the participating institutions. Interaction with the MTMT will be fully possible using a web interface (e.g. without using the client) - but this mode of usage is mainly intended for individual researchers. The client - using different plug-ins - will be able to connect to the local information systems of the universities, connect locally subscribed databases, and can hold site-specific data tables, enabling librarians to enter non-global information into the database. These clients will be capable (with the plug-ins) to produce statistics tailored for local needs, complemented with local information. Batch editing of institutional records will also be possible with the clients.

4 Challenges

As we have seen, the MTMT will ingest and integrate data from various sources. One of the challenges we face is the filtering of the duplicate entries. There is no single magic solution for this - comparison of meta-data is not highly effective, and full text is not available at MTMT, sometimes we do not even have links to it. The MTMT will provide tools for the authors or institutional librarians to find, mark and eventually combine duplicate entries. The most effective would be the use of globally unique identifiers as DOIs - but such identifiers are, unfortunately, not always available. The MTMT will promote the use of unique identifiers by the publishers, the repositories and the scientists.

Another challenge is author disambiguation. Here also the use of global unique identifiers will be the solution,

whenever ORCID [4] brings fruit. Some commercial databases already use author IDs. For the effective operation of the MTMT it will be necessary to convince Hungarian publishers to use unique document and author identifiers. It would be a great advantage if MTMT could receive meta-data from national scientific publishers, complete with those IDs. There is a similar demand towards the repositories: to store globally unique identifiers whenever available and to offer them to the MTMT harvester.

5 Data Policy

Harvesting will be confined to participating institutions and contracting partners. Public harvesting is not planned but the results of targeted queries could be downloaded for import elsewhere. The use of data for commercial purposes will not be permitted. Scientific or administrative use of large quantities of data will be by contract and personal data use will follow the wishes of the individuals involved.

6 Summary

As we have seen, the MTMT is designed to operate in the web of scientific information services, co-operating with other nodes using standard protocols and APIs. It uses a hybrid model for data ingest and curation: automatic harvest, expert control and the possibility of researcher intervention all play a role. Workflow is to be engineered in a way to minimise the workload of the scientist (and the librarian), not only for the MTMT, but in a broader sense: a publication entered into the web of interconnected services need not be handled by humans once more. In this way we hope to build a service which eases the burden of paperwork on the authors, while provides a solid foundation for science management.

Anyone will be able to query the MTMT using the web interface, but harvesting will be only possible for member organizations. Commercial re-use will not be permitted.

References

1. Hungarian National Scientific Bibliography (MTMT), <http://www.mtmt.hu/>
2. National Academic Research and Collaborations Information System (NARCIS), <http://www.narcis.nl/?Language=en>
3. Croatian Scientific Bibliography (CROSBIB), <http://crosbi.znanstvenici.hr/index.html?lang=EN>
4. Open Researcher and Contributor ID (ORCID), <http://orcid.org/>

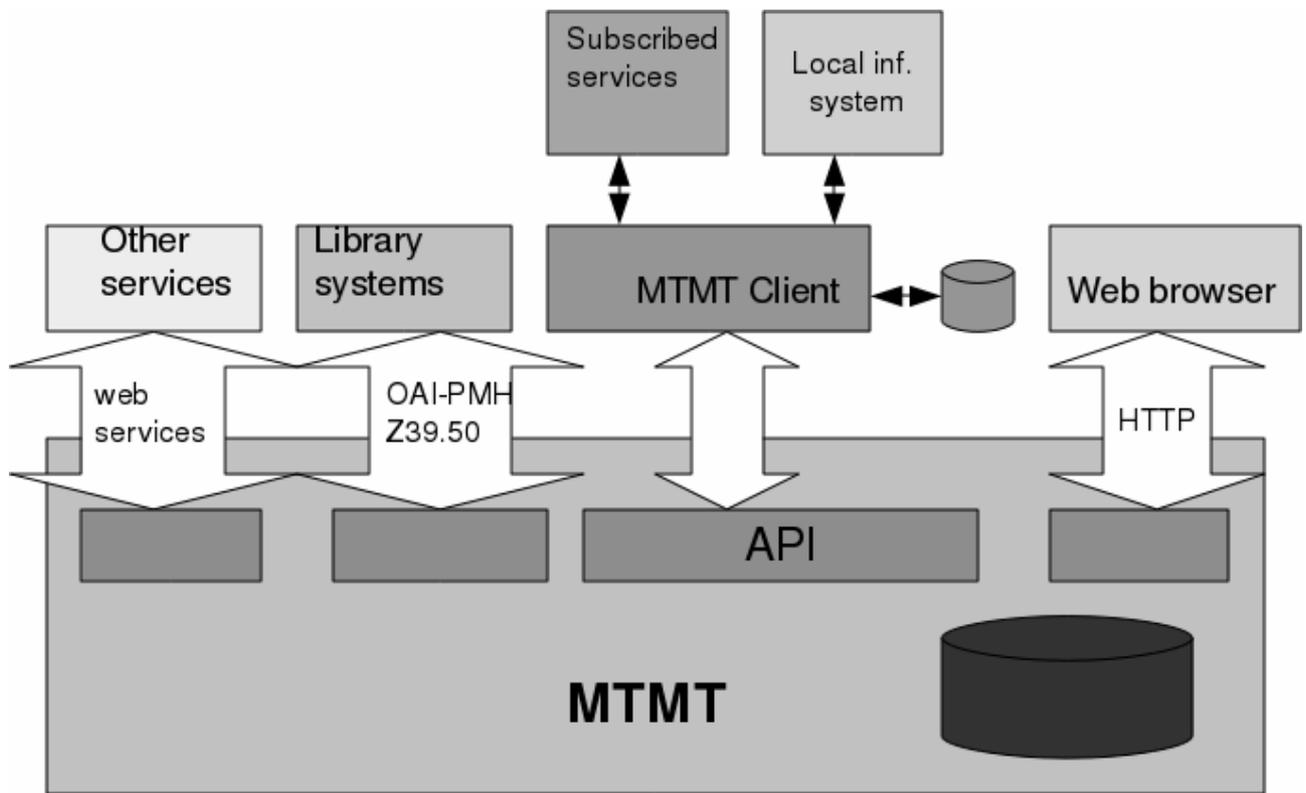


Fig.1. The architecture of the MTMT.