HunCRIS - towards semantic interoperability of CRIS-es

Adam Tichy-Racs
atichy-racs@omikk.bme.hu

Hungarian National Research Registry Unit
of the National Technical Information Centre and Library
at the Budapest University of Economics and Technology
(BME OMIKK)

Scope of presentation

Role of HunCRIS in the Hungarian National Innovation System

HunCRIS[1] is an information system of publicly financed Hungarian research and technology development projects, regardless the specific source of funding within the state budget. There are several sources of that sort

- regional and thematic programmes of National Development Plan using Cohesion Fund of the European Union;
- Fund for Research and Technological Innovation;
- Hungarian Scientific Research Fund (OTKA);
- Subsidy to Joint Research Units of the Hungarian Academy of Sciences and of the host institutions;
- Programmes for space research cooperation;
- Research of Lake Balaton;
- Social Studies of National Interest;
- Research programmes based on bilateral agreement on Scientific and Technological Cooperation;
- Programmes financed by ministries, etc.

HunCRIS was established in 2001 by a Government Resolution[2] in accordance with the association negotiations between Hungary and the European Union. The resolution took into account the Commission Recommendation of 6 May 1991 concerning the harmonization within the Community of research and technological development databases (91/337/CEE)[3]. The technical specifications for describing the projects are set out in Annex I. Today the specifications are known as the Common European Research Information Format (CERIF). There is „a multilingual and multidisciplinary common classification system for setting up the national databases referred to in point 1. This system must be flexible enough to allow regular adaptation to changing science and technology requirements. The classification system

[1] Available at https://nkr.info.omikk.bme.hu
[2] See: http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A0100160.KOR (in Hungarian)
[3] See: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31991H0337:EN:HTML

referred to is set out in Annex II". Furthermore, it is ready „to adopt a dynamic, multilingual and multidisciplinary common thesaurus (structured language)".

The declared goals of running HunCRIS are:

- Increase the transparency of using public money on research and technological innovation,
- Promote the use of research and development results,
- Supply proposal and project evaluation process with relevant and comparable information,
- Help finding consortium partners for national and multinational projects.

The software and the basic workflow in HunCRIS

The concept of HunCRIS software was set in 2001 and has been modified several times since then. The database behind is a CERIF-2000 compatible structure realized in ORACLE database management system. It has been upgraded to follow new releases of CERIF[4]. There is a JSP application supporting the workflow, and increasingly supporting user demands. There are additional applications to generate specific outputs human readable documents and files or computer readable images in CERIF compatible XML files. The CERIF compatibility of the metadata in these images enables **syntactic interoperability**.

Project data arrive one-by-one or in batch files. After a short formal check the major part of the data is entered into the database. The project specific keywords are added next. In case of earlier occurrence, that is an equivalent expression both in Hungarian and in English has been recorded, there is no repeated recording, but the earlier keyword is linked to the project description. If the keyword is new, it is entered in an unsorted list. Organizations and their units are linked from a structured list, their roles in the project, participation and project funding data are given. Although the top level is identified by tax number, there is no national identifier at lower level that needs some interaction. Finally the participating persons are linked to the project one by one manually, because we are not allowed to use any national identifier. The validity and uniqueness of data are checked before the data become public.

The unsorted list of keywords is regularly checked. First the formal properties of them are checked or corrected if needed. They should be scientific field specific being in noun phrase, single if plurality is not emphasized, the two lingual versions close to each other as much as possible, and so on. It is checked again, whether the modified Hungarian or English expressions have been registered earlier. If so, the other is changed to create a bilingual equivalent. Finally the elements in the unsorted list are inserted into their proper position in the ever-growing thesaurus. Since the upload of the original 2067 expressions of the EPSS thesaurus used to classify the project proposals in the Framework Programme, there are more than 13,500 new expressions have been added. These clones and nearly clones need regular elimination, so even equivalence in the list of the condensed, i.e. no space and no dash, expressions are checked.

There are regular checks of supposedly repeated recording of researchers, and there is also regular check of organizational structures.

When a record is made public, it is available for anybody immediately on HunCRIS site..

[4] See: http://www.eurocris.org/cerif/cerif-releases/

HunCRIS Services

HunCRIS is a bilingual information service that is available for everyone using any of the popular modern browsers. Selecting English one can make a simple registration or is immediately allowed to use the service without any sort of registration. For registered users there are more services provided and using the system becomes much comfortable.

Projects, Organizations/Units or Researchers can be searched using different fields and their Boolean combinations in two levels. (For multilevel expressions you should store and reuse simpler queries.)

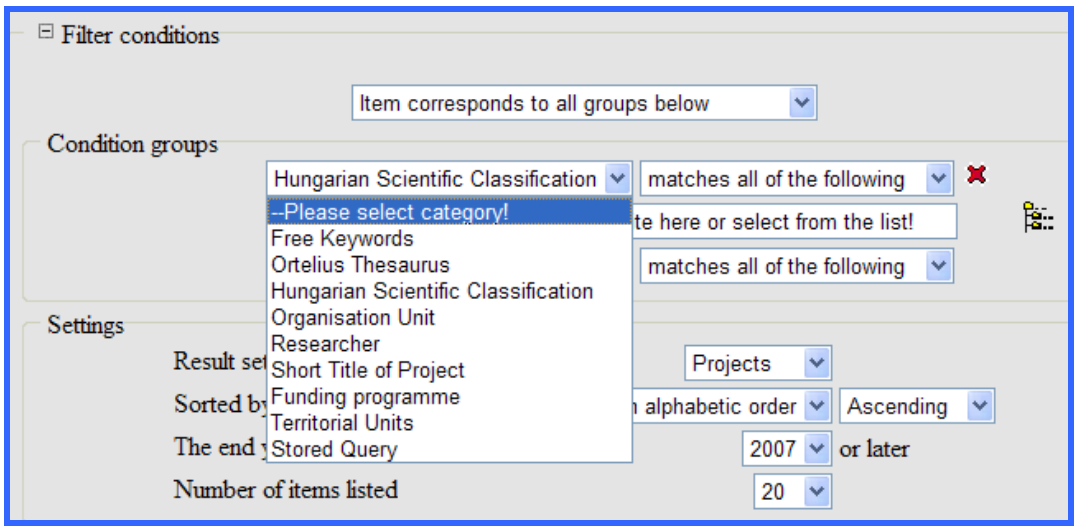


Figure 1. Query categories in HunCRIS

The results can be filtered according to the closing year of the project, where the default is the second year before current one. Registered users can store and reuse their queries, and they can generate simpler hyperlinks according to their queries, which results the query and result list always with the current data at the moment of recall.

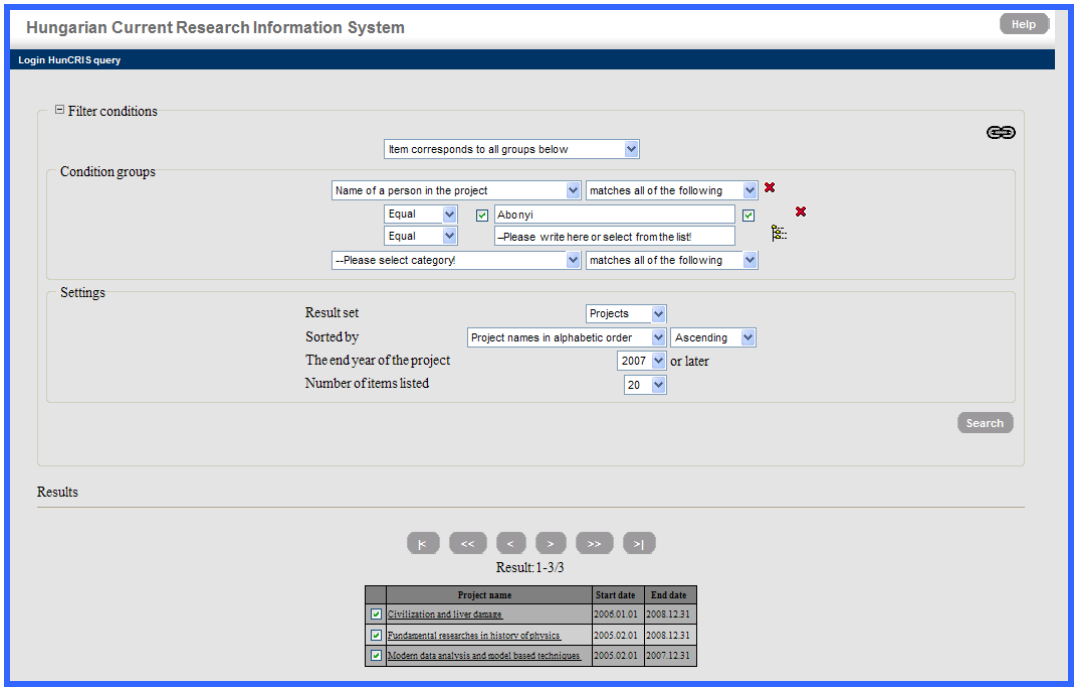


Figure 2. Query and result page

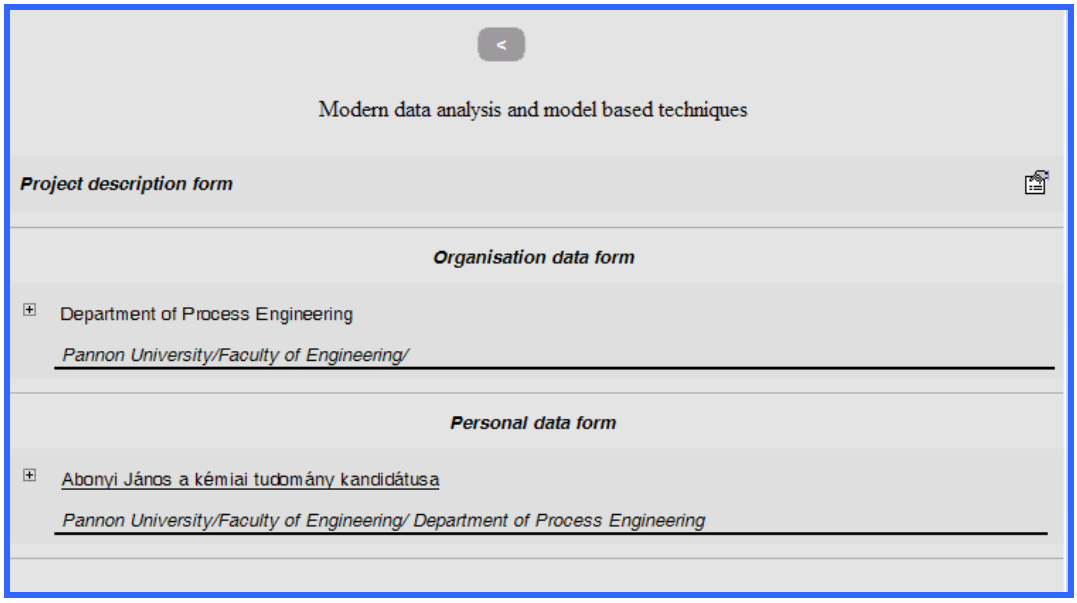


Figure 3. General project information: Title. Organization(s), Researcher(s)

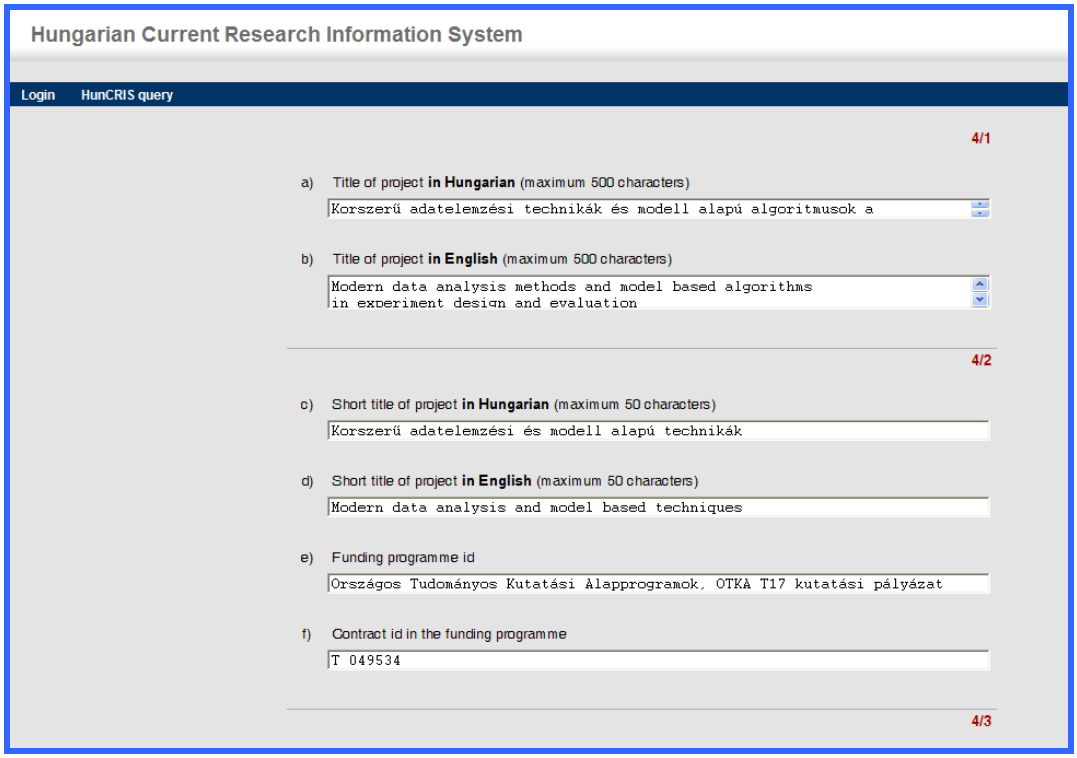
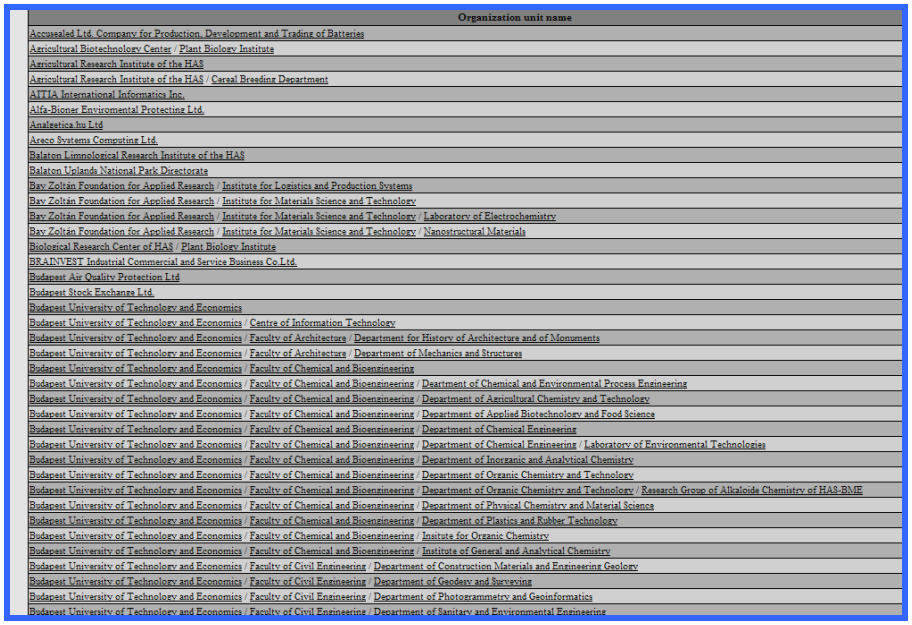


Figure 4. Top of project description page

The query result can be changed from Projects to the participating Organizations/Units or researchers, and clicking on any item the list of its projects is resulted.



Organization unit name
Accusealed Ltd. Company for Production, Development and Trading of Batteries
Agricultural Biotechnology Center / Plant Biology Institute
Agricultural Research Institute of the HAS
Agricultural Research Institute of the HAS / Cereal Breeding Department
ATTIA International Informatics Inc.
Alfa-Bioner Enviromental Protecting Ltd.
Analextica hu Ltd
Areco Systems Computing Ltd.
Balaton Limnological Research Institute of the HAS
Balaton Uplands National Park Directorate
Bay Zoltán Foundation for Applied Research / Institute for Logistics and Production Systems
Bay Zoltán Foundation for Applied Research / Institute for Materials Science and Technology
Bay Zoltán Foundation for Applied Research / Institute for Materials Science and Technology / Laboratory of Electrochemistry
Bay Zoltán Foundation for Applied Research / Institute for Materials Science and Technology / Nanostructural Materials
Biological Research Center of HAS / Plant Biology Institute
BRAINVEST Industrial Commercial and Service Business Co.Ltd.
Budapest Air Quality Protection Ltd
Budapest Stock Exchange Ltd.
Budapest University of Technology and Economics
Budapest University of Technology and Economics / Centre of Information Technology
Budapest University of Technology and Economics / Faculty of Architecture / Department for History of Architecture and of Monuments
Budapest University of Technology and Economics / Faculty of Architecture / Department of Mechanics and Structures
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Chemical and Environmental Process Engineering
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Agricultural Chemistry and Technology
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Applied Biotechnology and Food Science
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Chemical Engineering
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Chemical Engineering / Laboratory of Environmental Technologies
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Inorganic and Analytical Chemistry
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Organic Chemistry and Technology
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Organic Chemistry and Technology / Research Group of Alkaloids Chemistry of HAS-BME
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Physical Chemistry and Material Science
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Department of Plastics and Rubber Technology
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Institute for Organic Chemistry
Budapest University of Technology and Economics / Faculty of Chemical and Bioengineering / Institute of General and Analytical Chemistry
Budapest University of Technology and Economics / Faculty of Civil Engineering / Department of Construction Materials and Engineering Geology
Budapest University of Technology and Economics / Faculty of Civil Engineering / Department of Geodesy and Surveying
Budapest University of Technology and Economics / Faculty of Civil Engineering / Department of Photogrammetry and Geoinformatics
Budapest University of Technology and Economics / Faculty of Civil Engineering / Department of Sanitary and Environmental Engineering

Figure 5. Participating units and partners in the projects of Budapest University of Technology and Economics (BME)

Result: 1-200/864

Ábrahám György
Ács Péterné
Ádám József
Ádány Sándor
Ásai Béla
Agod Attila Péter
Almássy Zsuzsa
Andor György
Anna Péter
Antal Péter
Antal Péter
Aparvi Barnabás
Aradi Petra
Arató Péter
Babarczy Anna
Bazi Katalin
Bajor Gábor
Bakó Péter
Balázs Andrea
Balázs Barbara
Balázs László
Balázs Márton
Balázs Tibor
Bálint István
Bálint Péter
Ballagi Áron
Balogh Attila
Balogh Attila
Balogh Edina
Balogh László
Balogh Tibor
Bánk András
Bánky Tamás

Figure 6. Researchers in the projects of BME

https://nkr.info.omikk.bme.hu/ff004/show.do?filtermode=and&group_id=1&group_type=P&conditions_and=and&condition_id=19683&condition_eq=eq&condition_value=+Bagi+Katalin&condition_group_id=1&classname=projectSearch

Figure 7. Code of hyperlink from researcher's list (in Figure 6) to the projects of researcher Bagi Katalin

Following the hyperlink of the lists in Figure 5 or 6 the original query conditions are not inherited, that is projects out of the previous query conditions are found, too. To keep the original query and filter to the list item, one has to insert the selected item into the previous query. Storing queries is one of the privileges of registered users.

Registration as an explorer is simple and automatic. To get higher authorization, i.e. analyst or data provider, needs additional action from operators and in some cases a special permission from the National Office for Research and Technology (NKTH).

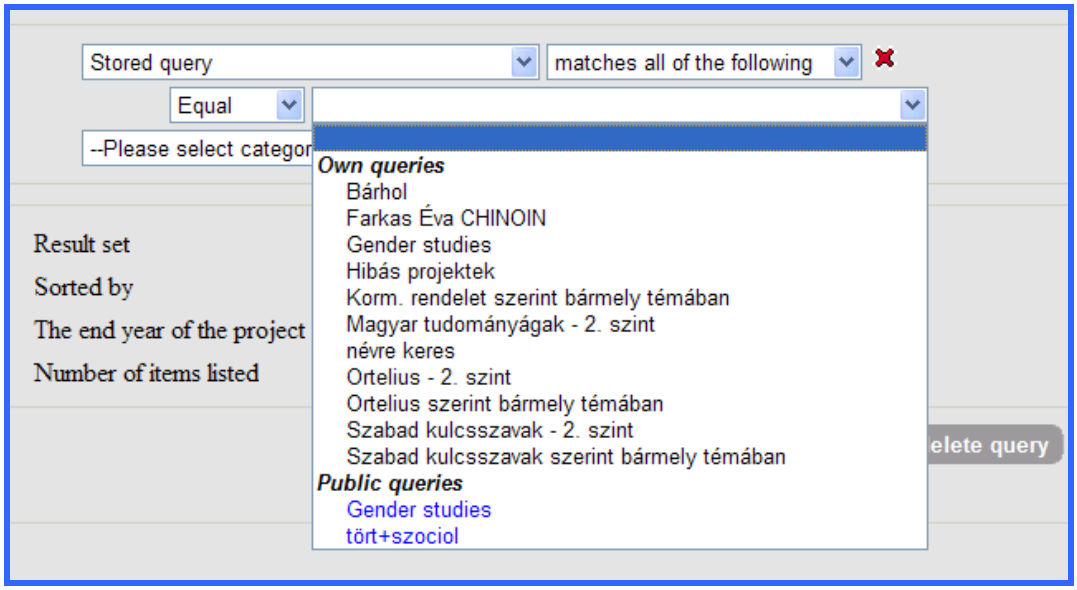


Figure 8. Reusing stored queries

Thesauri and controlled terms in HunCRIS

To help cooperation with special national services and international cooperation there are different sets of controlled terms in HunCRIS.

There are **three sets of academic fields**;

- A structured set of **free keywords** provided by project managers, that is arranged into thesaurus-like (!) structure by HunCRIS team;
- **Ortelius Thesaurus** is the successor of the classification system given in Annex II. of the 91/337/CEE recommendation. Since its first installation in HunCRIS it has been extended by the expressions used in EPSS of the European Commission;
- the official **national classification** system of scientific fields and branches.

Funding Programmes, Calls and Actions are arranged in tree structure.
Academic titles are arranged as general terms and below the top level terms there are more specific ones to indicate academic fields.
Qualifications are arranged according to their levels and fields.
Project positions of researchers
Work positions, that is positions in the researcher's organisation are also controlled terms
Territorial Units is a tree structure of NUTS-2 and NUTS-3 regions in Hungary, their microregions, settlements and even parts of settlements.

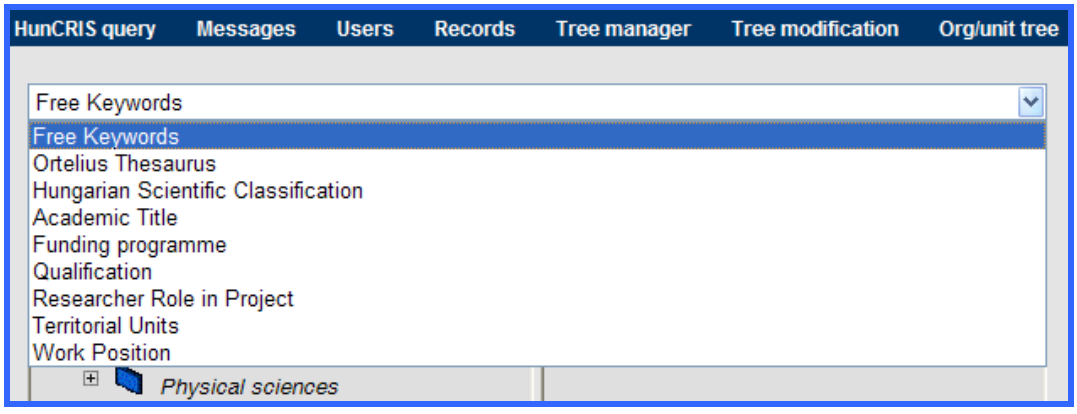


Figure 9. List of thesauri and controlled terms

The sets of controlled terms can be modified or extended, and new sets can be added if it looks reasonable.

Semantic features of HunCRIS

While syntactic interoperability is based on metadata, semantic interoperability depends on the quality of controlled terms. Projects are primarily indexed by the project manager, and that indexing is revised by HunCRIS team. These primary indexes are indicated in bold face in Figure 10. (It is the same project that was found in Figures 2 to 4.) Indexing terms in normal face are derived from the structured list of terms, adding all items above the indexing terms. (The reasons of alphabetic order will be explained later!)

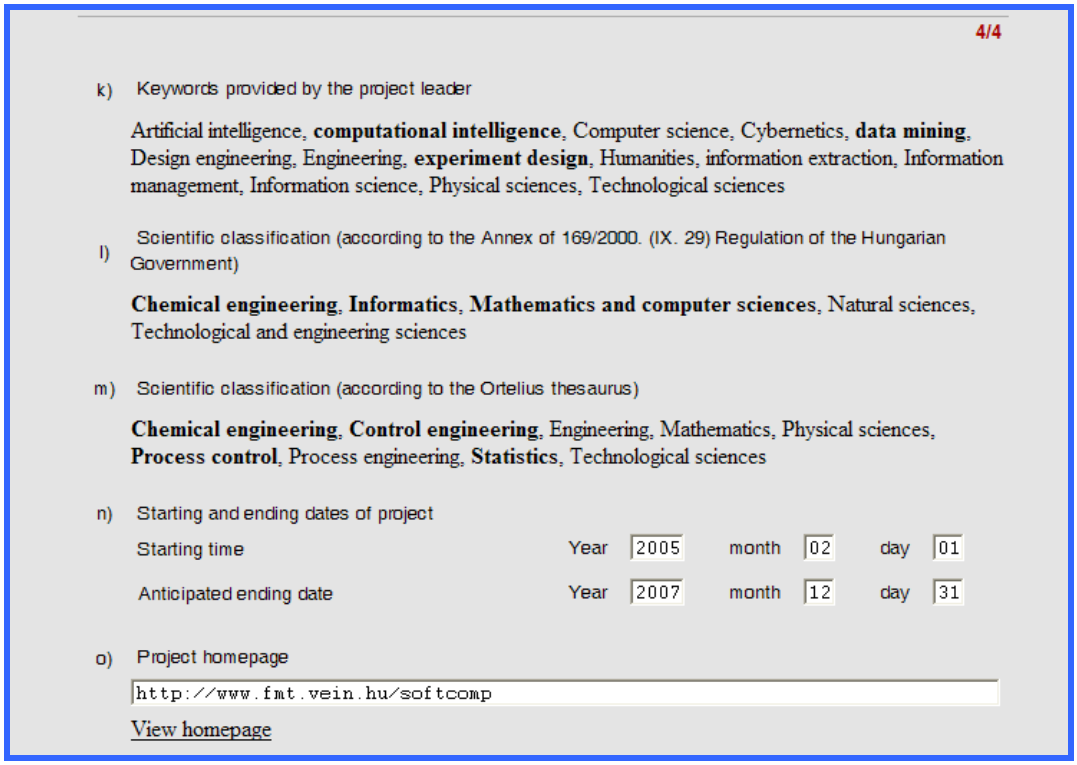


Figure 10. Bottom of project description page

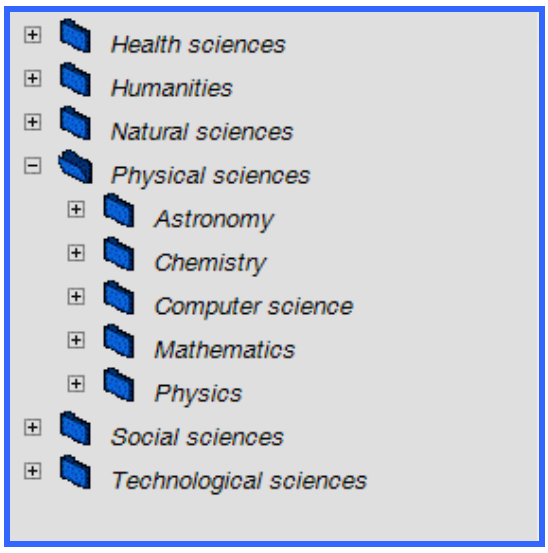


Figure 11. High level items in tree

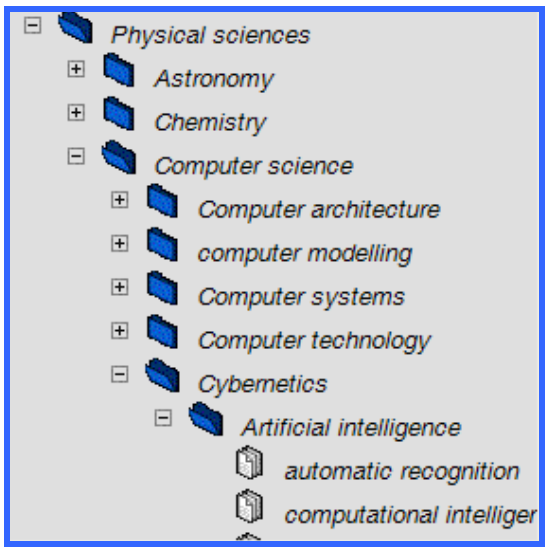


Figure 12. Lower level items in tree

Maintenance of the structure means, first of all, inserting all new expressions to the proper place of tree. Some items should be found by different approaches. To help finding these expressions, more than one "parent" can be marked, that is the item can be moved to a given position and can be copied to another, as it can be seen in Figure 13. (All new items are put into an unsorted set, and will be moved to their position manually.)

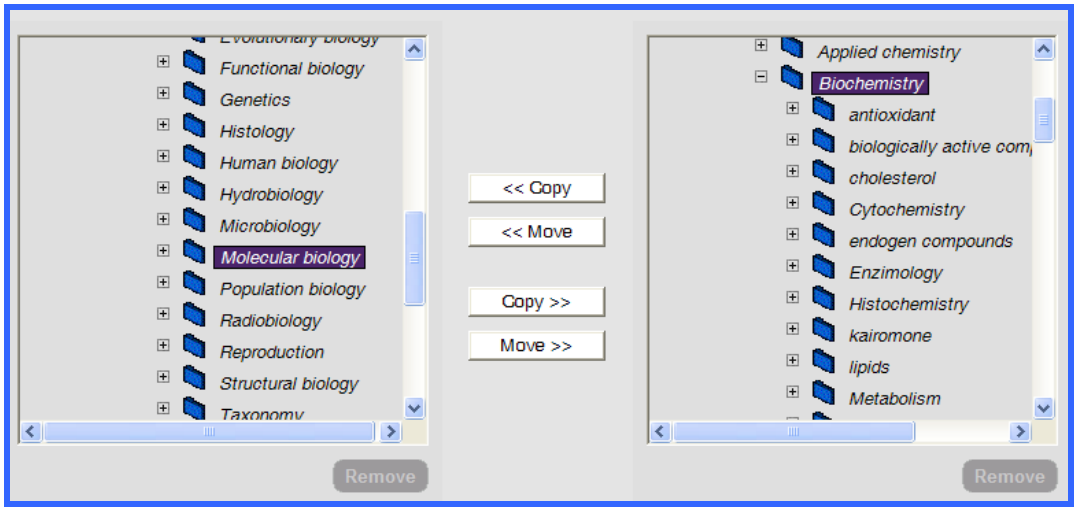


Figure 13. Tree manager with Move Copy and Remove functions that allow us to include one expression in multiple categories (parents)

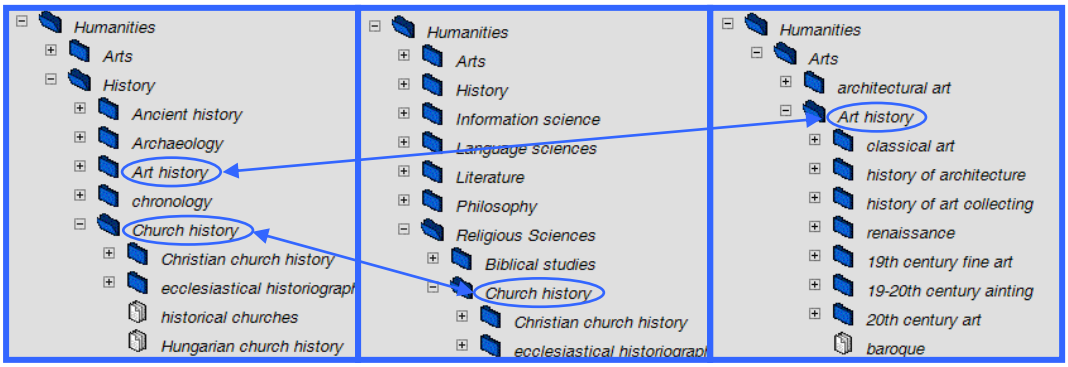


Figure 14. "Art History" and "Church history" can be found as parts of "History", and as parts of "Religious sciences" and of "Arts", respectively

Towards semantic interoperability

Matching metadata structures of CRIS systems have been set to reach syntactic interoperability. For semantic interoperability we have to use the same vocabularies[5], which is a much tougher problem. All of us use this or that sort of thesaurus or classification system or however we call them. In HunCRIS we have to use a two level national classification system and we have installed one of the thesauri of the European Commission, the one which is used in the Electronic Proposal Submission System (EPSS), but we have not installed Fields of Science and Technology for Frascati Manual (FoS), which is recommended by OECD for R&D statistics. Elaboration of any of the above mentioned structured vocabularies needed several years, and, as you see, there are parallel and highly incompatible vocabularies.

[5] See: http://www.eurocris.org/fileadmin/cerif-2008/CERIF2008_1.0_Semantics.pdf

Reading heavy books on science philosophy one has to realize that the meaning of the same expression is different from community to community, their meaning varies with the purpose of use, so we are not able to develop a single universal system for all CRIS-es.

Our proposal is to use the local vocabularies together in an interoperable way: should the XML representation of project data contain not just the most *precise indexes* that are added to the project description by the project manager, but use *extended indexing*, that is all the expressions above them in the given structured set of expressions, as it can be seen in the figure below:

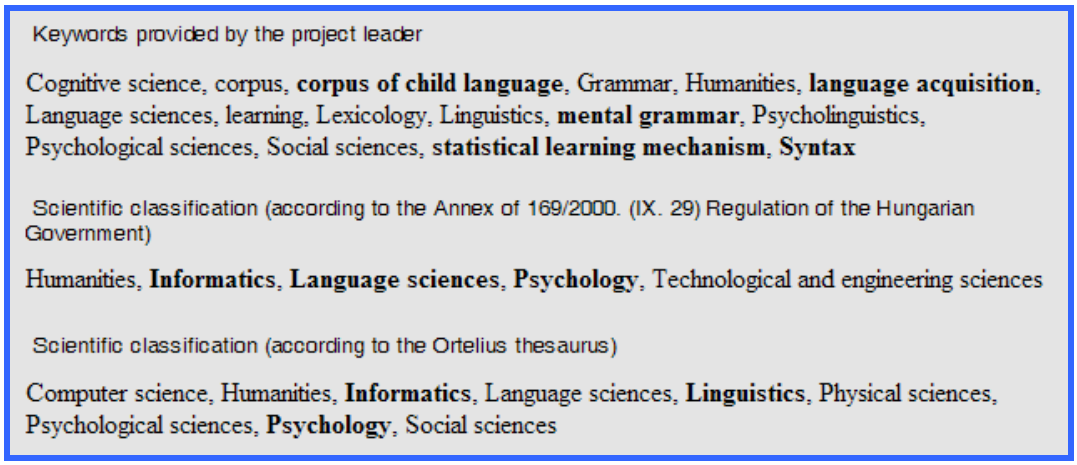


Figure 15. Independent index sets of the same project

To demonstrate the interoperability of different vocabularies we compare the three sets of indexes in Figure 15 that describe our randomly selected project. The common terms of the different sets are collected in Table 1.

Index sets 1 and 2	Index sets 1 and 3	Index sets 2 and 3	Index sets 1, 2 and 3
Humanities Language sciences	Humanities Language sciences Linguistics Psychological sciences Social sciences	Humanities Informatics Language sciences Psychology	Humanities Language sciences

Table 1. Common elements of index sets in the extended index set of a selected project

If we suppose, that this specific project is stored in three different CRIS-es using Vocabulary-1 Vocabulary-2 and Vocabulary-3, we can realize, that their descriptions are not too far from each other, but there are meaningful differences, too. This means, that any pairs of the three systems are interoperable! It is obvious, that the quality of interoperated systems is degraded much more, if project data are harvested from System-2 to System-1 and then read by System-3, than in case of direct reading from System-1 to System-3, but there is still some level of semantic interoperability.

The quality of common data provided through portal will depend on the level of semantic interoperability of systems. All information systems are better focused on their own information than on information of any other information systems, so portals using their own structured sets of expression are less focused to other systems.

Recommendations

There are several CRIS-es all over the European Union and beyond. HunCRIS is just one of them. These systems are interoperable syntactically, or at least their data can be converted into syntactically interoperable XML formats with the same metadata, or using metadata library. These converted data should include all terms of extended indexes of projects.

The basic property of any CRIS portal is to collect metadata, and support common use of data stored in separate CRIS-es. For higher level use of data these CRIS-es have to use controlled terms, but it is more difficult problem than using the same XML schema. For this reason, the portal should use the terms of a selected system, favourably the local one. Since then, the level of interoperability depends on the level of similarity of controlled terms in the selected systems.

The portal should provide more added value than just being a common access point to different systems. It should be equipped with advanced visualization, analytic and semantic tools, such as identifying cooperating networks, uncover multiple financing, search for similar activities, knowledge mapping and trend analysis.

As a consequence of different thematic orientations of local systems, the local semantic resources will be different. For example, a portal of research projects in social sciences, another one in biodiversity and a third in information society technologies can be perfectly similar in their metadata structure, but they have extremely small similarities in their information content. For that reason, if the new portal will be installed at any specified CRIS, it cannot offer its added value to other portals completely. To gain optimal result, we should install the new instrument to all participating CRIS-es.

There are additional advantages of using multiple entry points to cooperating systems. The new portal can be customized to improve the service locally, and the results can be shared with the partners. To improve the services the controlled terms can converge smoothly. Last, but not least, the continuity and the quality of service will not depend on the changing conditions of a specified service provider, so this solution will ensure sustainability of the service.