

A National CRIS Infrastructure as the Cornerstone of Transparency in the Research Domain

Tomáš Chudlarský, Jan Dvořák

Institute of Information Studies and Librarianship,
Faculty of Arts, Charles University in Prague, Czech Republic

Summary

The Czech Research, Development and Innovation Information System is the key component of the national research information infrastructure in the Czech Republic. The role of the national CRIS has traditionally been strong. We present the usage data of the Czech R&D&I Information System: it is being used massively and frequently. We thus argue that for the trust-challenged niche of the Czech research domain, the information system presents a real cornerstone of transparency.

1 Introduction

Research that is funded from the budget of a national state calls for being documented in a national Current Research Information System (CRIS). Such a CRIS concentrates valuable research information that is potentially very useful. The question arises whether this potential is actually being realized.

Trust in public institutions seems to be below the OECD average in the Czech Republic. The Synthesis Report of the International Audit of Research, Development & Innovation in the Czech Republic project gives the following first and foremost recommendation: *Building trust in government and among members of the R&D&I community is a precondition for successful performance of the National Research and Innovation System. Such trust is based on fair principles transparently and impartially implemented combined with conspicuous punishment for those at any level who break the rules.* (Arnold 2011, p. 55).

Open and accessible research information contributes to transparency in the research domain (e.g. Demšar, Boh 2008). Transparency leads to an enhanced level of trust, more open competition, strengthened equality of opportunities, and information access equality.

Current Research Information Systems are built on all levels in the research domain. The usefulness of a university or a research institute's CRIS is verified every day by the staff, students and collaborators. A funding agency's CRIS is essential for keeping track of the agency's assets and liabilities. But it is not clear if this scales up to the national level – national CRIS fulfil a different set of roles (Alroe, Rasmussen 2010). The wide usage of a national CRIS, albeit expected, is not straightforward. Only when the CRIS is actively being used, it is defensible to speak about its contribution to transparency.

In Section 2 we provide a brief description of the Czech Research, Development and Innovation Information System, a fundamental component of the Czech national research information infrastructure. In Section 3 we present its typical users and usage statistics. Then we draw conclusions (Section 4).

2 National CRIS Infrastructure in the Czech Republic

The national CRIS infrastructure consists of the following components: the central database with the web presentation of its contents, and the data collection system. The latter includes a publicly available data integrity checking service and a set of tools (“Vklap”) to support the data collection process.

A previous state of the information system was described in (Dvořák, Souček 2008). There has been development since that time.

2.1 The central database

The information in the central database is kept in the following IS parts:

- Registry of research projects (CEP) – collects information about research projects that were at least partly funded from public sources. This is the oldest part, its history reaches back to 1994.
- Registry of institutional research plans¹ (CEZ) – very similar in structure to CEP.
- Registry of results (RIV) – collects information about research outputs (publications, patents, technologies).
- Registry of R&D&I calls (VES).
- Registry of other R&D&I activities (CEA) – one part of it tracks funding schemes.
- Tools to support the preparation of the R&D&I state budget proposal (SR) – no public interface to the data here.

Table 1 gives an idea about the numbers of objects of various kinds in the information system. The data collected needs to be integrated with similar data collected in previous years, with data from other parts of the information system, as well as with external data (from the statistics office, the national library, ISBN International, CORDIS², etc.). All these operations improve the information value and the quality of the data.

Table 1: Numbers of main objects in the Czech R&D&I IS and their annual increments

Objects	Acronym	Object count	Increase per year	Collected since
Research projects	CEP	37,000	1,800	1994
Institutional research plans	CEZ	889	0	1998
R&D result records	RIV	870,000	100,000	1998
Cleansed R&D results		630,000	60,000	
Calls	VES	458	30	2000
Funding schemes	CEA	235	10	2010 (previously kept in a master table)

¹ Also known as Research Intentions - a former vehicle of institutional funding

² FP6 and FP7 calls are sourced from CORDIS

2.2 The web presentation of the IS data

The data of the information system is made available on the web³, with the exception of classified projects and their results, of trade secrets, and of personal data (identifier, personal address, personal contact information). The web presentation is the main channel of transparency.

The presentation is in two languages: Czech and English. Certain data about results (such as title, abstract) is in the original language.

The web presentation of the data allows user searches on the following entities: Funders, Programmes, Organizations, Calls, Projects, Institutional Research Plans, and Results. These are also the entities for which detailed pages are produced. Context is provided through hypertext links between the detail pages.

The data of found objects (up to 50,000 rows) can be exported in ZIP archives containing one of the two available formats: XHTML (can be opened in MS Excel) and DBF.

2.3 The data collection system

Since its inception, the Czech R&D&I Information System uses a batch-oriented data collection system. E.g. instead of sending the information about every research project individually, a funder sends a batch of data every year on *all* on-going projects in a research programme. If something changes or needs to be corrected, the funder sends a new version of the batch that replaces the original one in the database. Large batches can contain information about hundreds, sometimes thousands of objects.

This corresponds to the character of the information system: rather than an on-line transaction system, it's a decision support one. The batch data collection has the following three major advantages over interactive filling-out of data:

1. Clear responsibility of parties.
2. Improved data quality, especially its completeness aspect.
3. Greater resilience against data loss.

In the history of the Information System, an interactive component for the data collection was attempted. It turned out that the interactive approach does not scale up to the national level: the main reason was the complexity and great variety of approval processes at all levels. The experience so far suggests that the interactive data collection of research result metainformation works up to the level of a university (or the Academy of Sciences as a whole), but not beyond.

To support the data collection process there is the universally accessible service to validate a data batch. Its usage is mandatory before delivering the data batch to the central database. This explicit separation of interface makes it possible to deliver data from heterogeneous CRIS of the funders and of the research-performing organizations.

The information system also provides a default set of tools - the "Vklap" - to support the data collection process. In fact, this can be used as a stand-alone CRIS for smaller data sets (e.g. a smaller research institute or a company).

³ www.isvav.cz/locale/en_US/

2.4 Division of responsibilities

As a general rule, the flow of information copies the financial flow in reversed direction: every participant gives an account of how they spend the funding.

Funders report to the Czech R&D&I Information System on the research activities they fund: research projects, institutional research plans, and other kinds of research activities. Here the responsibility is clear as each research activity has exactly one state funder.

With research results (outputs) the situation is more complex. Results are created by individual researchers while working on a research activity at a research-performing organization: these organizations report the results to the funders of the research activities. The funders verify the existence of the results and their relevance to the indicated research activities, and forward the data batch to the R&D&I Information System. Results often arise from cooperation of several authors from potentially multiple institutions where they work on potentially multiple research activities with potentially multiple funders. As a consequence, the information about one result may travel by several paths. Each of these paths represents the result from a specific perspective connected with a particular responsibility setting. One of the core functionalities on the central level is the consolidation of these pieces of information into a coherent picture.

3 Czech R&D&I Information System Users, Usage and Statistics

The Czech Research, Development and Innovation Information System has a wide audience. The primary user is the keeper of the information system: the Research, Development and Innovation Council. This is an advisory body to Government itself.

The funders (ministries and the funding agencies) are also frequent users of the information system: theirs is the responsibility for the data in it. However, their second type of usage is in the evaluation of project proposals: the reviewers and the panels need to see the context of previous research done by the funding applicants.

The research-performing organizations (universities, institutes of the Academy of Sciences of the Czech Republic, sectorial research institutes, business enterprise development divisions) also make heavy use of the information system. They supply the metadata about their outputs, which make the basis of the evaluation. They often search for information about other organizations in the same field of science, about their potential collaborators, and competition.

Innovative businesses draw information about current research topics and outputs from the CRIS.

In the end, all these players in the government-funded research and development arena need to defend “before the tax-payer” the efficiency of their spending or distributing the funding.

The typical usages of the information from a national CRIS include:

- The public (including researchers themselves) can access information about research activities, their funding, the persons and organizations that participate in them, and about their results.
- Government bodies and agencies can compare the research results of organizations, organization units, research teams and individual researchers, along with funding.
- Project proposal evaluation has to take the context into account: prior research activities, existing expertise of the project team, research capacities of the organizations.

- Research outputs are visible and concentrated in one place. The added value is that one can see intra-national collaborations.
- Project outputs can be confronted with project inputs.
- A directory of research performing institutions of a country, possibly with subject profiles.
- Justification of the expenditures to the tax payer.

We will study the period from January 2010 until April 2012 (inclusive).

In this period the system has served about 152 million raw HTTP requests. Out of this number, 78 million requests actually asked for data or some action, the rest were images and other auxiliary files.

Most of the load was generated by web search engines: Google alone accounts for about one half of the requests. We identified only 8.9 million requests that are not related to web search engines.

6.7 million requests originate from hosts in the “.cz” internet domain. The domain could not be determined for more than 1.4 million requests.

Out of these 6.7 million requests, $\frac{2}{3}$ seem to be produced by internet browsers and $\frac{1}{3}$ by download tools or agents designed specifically to crawl the Czech R&D&I IS web - these tools are run by universities and the Academy of Sciences.

Our study aims to answer the following three basic questions:

- How the CRIS usage level changes in time;
- Who uses the CRIS;
- What are the most frequently used parts of the CRIS.

The monthly usage from beginning of year 2010 to April 2012 is described in the chart on Figure 1. The chart conveys the following facts:

- The overall usage levels of the CRIS grow gradually.
- Periods of high load are observed following the release of the evaluation results (March 1, 2010; February 7, 2011; January 27, 2012).
- Periods of low load are observed in the months outside the most important data collection (spring).

Also, we interpret the variance in the load throughout the year as an evidence that we have sufficiently cleansed the data of non-specific machine traffic (especially the web search engines).

We further wanted to identify the request originators. We translated the IP addresses of the hosts that placed the requests, to domain names. We only consider the national “.cz” domain. The results are presented in Table 1.

It is verified that the R&D&I Council is not the only user of the Czech R&D&I IS. It is massively used from all universities and the Academy of Sciences. The research-performing institutions are the biggest user. But also ministries and other funding bodies use it.

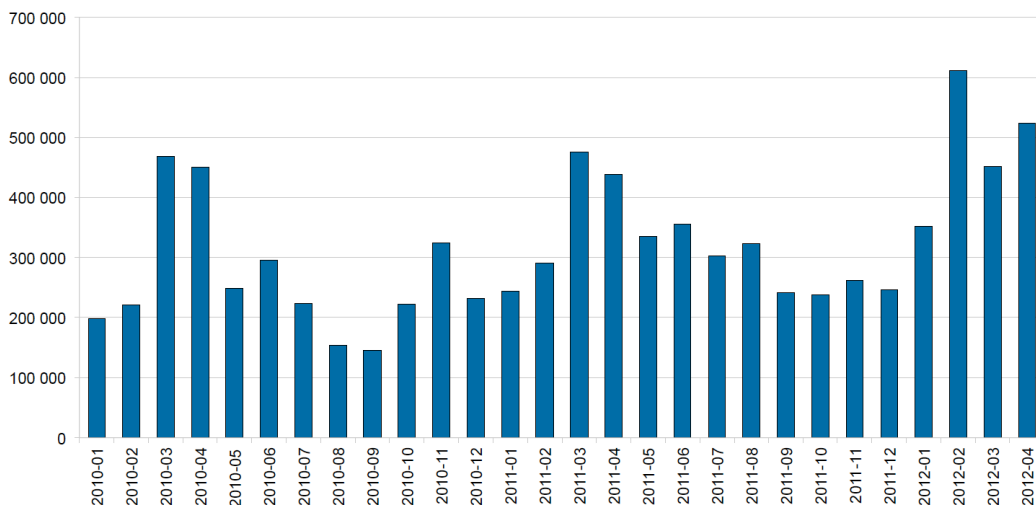


Figure 1: Requests on the Czech R&D&I IS by months (not including the web search engines; approx. 8.9 million pages total)

The number of unique IP addresses is an interesting piece of data. This number provides a reasonable lower estimate of the number of users of the CRIS in the institution. The academic institutions from the biggest 50 users sum up to more than 39 thousand unique IP addresses. There may be tens of automated agents in these institutions, so we conclude there are at least 39 thousand researchers and other staff at the universities and the Academy of Sciences. And this is just an estimate from below, there may be several people or machines behind one IP address.

Contrary to this, the load from funding bodies or the R&D&I Council Secretariat are usually concentrated to just one or few distinct IP addresses. This is due to the Internet security policies at the institutions, they often use a central Internet gateway.

These results are expected. Some partial facts that may look as surprises at first:

- The really large (14%) load generated by the Office of the Government. Most of this load is owed to internal tools that let the staff of the R&D&I Council Secretariat perform the tasks of importing in-coming data batches and controlling the data processing in the central database. These tools generate 11%. The remaining 3% are produced by the ad-hoc queries by the staff.
- The relatively low usage of the data by the Grant Agency and the Technology Agency. While the latter is a relatively new funder that has not yet gone through the whole cycle, the reasons for the low usage by the Grant Agency are not known. In both cases it can be guessed that their reviewers and evaluators use the CRIS from their institutions or homes.

Table 2: Requests on the Czech R&D&I IS (only from the “.cz” domain, since January 2010 to April 2012 ~ about 6.7 million requests).

Rank	Domain (reversed)	Hits	Hits (relative)	Distinct IP addresses	Institution
1	cz.vlada	952242	14,20%	2	Office of the Government
2	cz.cas	622134	9,28%	3448	Academy of Sciences
3	cz.iol	497092	7,41%	87685	a major internet provider for households
4	cz.cuni	472191	7,04%	5450	Charles University in Prague
5	cz.upcbroadband	372998	5,56%	67884	a major internet provider for households
6	cz.muni	345608	5,16%	4272	Masaryk University in Brno
7	cz.vsb	321163	4,79%	3639	Technical University Ostrava
8	cz.cvut	303298	4,52%	4260	Czech Technical University, Prague
9	cz.infoscience	133046	1,98%	4	IS maintainer
10	cz.o2isp	126982	1,89%	4268	a major internet provider for households
11	cz.tul	91930	1,37%	1916	Technical University Liberec
12	cz.vutbr	86902	1,30%	3831	Technical University Brno
13	cz.tmcz	74356	1,11%	11842	a major mobile internet provider
14	cz.zcu	71060	1,06%	1874	University of West Bohemia, Plzeň
15	cz.jcu	56083	0,84%	1822	University of South Bohemia
16	cz.osu	51325	0,77%	662	University of Ostrava
17	cz.eurotel	50788	0,76%	12663	a major mobile internet provider
18	cz.mendelu	48887	0,73%	1088	Mendel University Brno (agriculture)
19	cz.mkcr	47669	0,71%	1	Ministry of Culture
20	cz.env	46514	0,69%	2	Ministry of Environment
21	cz.upce	42968	0,64%	1690	University of Pardubice
22	cz.vscht	40946	0,61%	1144	Institute of Chemical Technology Prague
23	cz.ujep	35980	0,54%	452	University of Ústí nad Labem
24	cz.mvcr	35522	0,53%	2	Ministry of Interior
25	cz.slu	33490	0,50%	553	University of Silesia
26	cz.vzlu	29119	0,43%	1	Aerospace Research and Test Establishment
...					
34	cz.gov	21900	0,33%	13	Public Administration Portal
35	cz.msmt	21877	0,33%	1	Ministry of Education, Youth and Sports
...					
45	cz.army	15898	0,24%	4	Ministry of Defense
...					
48	cz.mpsv	15455	0,23%	6	Ministry of Labor and Social Affairs
...					
51	cz.nm	14584	0,22%	1	National Museum
...					
53	cz.nkp	14341	0,21%	7	National Library
...					
56	cz.czso	12967	0,19%	1	Czech Statistics Office

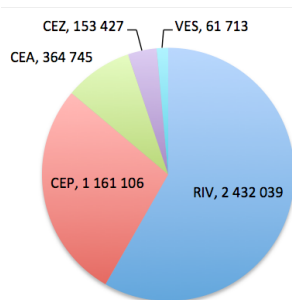


Figure 2: requests from the ".cz" hosts by type of objects accessed

Figure 2 shows that the results registry ("RIV") is the most frequently accessed component, whereas research projects ("CEP") come second. There is no doubt these two areas are among the most important ones for the transparency. Research project information allows the research funding flows to be analysed. When taken relatively to the number of objects of the respective kind, the research project information is accessed more frequently on the average than research results.

It can also be derived that the "Vklap" (the default tool to support the data collection process) is used frequently, and by many users. We have seen more than 73 thousand executions of this tool.

The publicly available web checking service was used almost 74 thousand times, while the batch ticket was created 22.5 thousand times. These figures support the importance of this validation service. The statistics also show about 19 thousand of data downloads through the export functionality of the web presentation of the R&D&I Information System data.

Let's note that the total load on the national CRIS infrastructure is much higher: 89% of requests for HTML pages are generated by web search engines. On top of that, auxiliary files such as images and styles are to be served too.

4 Conclusions

We have shown that the data of the Czech R&D&I Information System is accessed massively and frequently. There are about 300,000 requests per month (excluding web search engines). We have identified accesses by at least 39,000 academic users. This should be confronted with the approximately 78,000 of R&D personnel (head count) in the country, about 43,000 of them being researchers (Czech Statistics Office 2011). Also the funding bodies and the Research, Development and Innovation Council itself make frequent use of the information system. The usage is rising.

The Czech R&D&I Information System thus presents a real cornerstone for transparency in the research domain. Transparency of processes, which is called for in the recommendation of the International Audit project (Arnold 2011), can only be built on transparent data. We demonstrated that this transparency is achieved.

References

- Alroe B., Rasmussen H. S. (2010): National and institutional levels. Contribution at the euroCRIS membership meeting, Prague, November 8th-9th, 2010. Available from http://www.eurocris.org/Uploads/Web%20pages/members_meetings/201011%20-%20Prague,%20Czech%20Republic/Research%20Information%20Management%20-%20Bo%20Alroe%20&%20Henrik%20Steen%20Rasmussen.pdf [retrieved 15-Nov-2010].
- Arnold E. (2011): International Audit of Research, Development & Innovation in the Czech Republic - Synthesis Report. Technopolis Group 2011. Available from <http://audit-vav.reformy-msmt.cz/download-document/58-synteticka-zaverecna-zprava-aj> [retrieved 1-Dec-2011].

Czech Statistics Office (2011). *Research and Development Indicators 2010*. Available from http://czso.cz/csu/2011edicniplan.nsf/engkapitola/9601-11-eng_r_2011-0303 [retrieved 2-May-2012].

Demšar F., Boh T. (2008): Information systems as tools for managing research activities - how it works at the Slovenian Research Agency. In: *Get the Good CRIS Going: Ensuring Quality of Service for the User in the ERA*, Aleš Bošnjak, Maxmilian Stempfhuber (eds.), CRIS2008, proceedings from the 9th International Conference on Current Research Information Systems, Maribor, Slovenia, June 5th–7th, 2008, IZUM, ISBN 978-961-6133-38-8, pp. 11-19.

Dvořák J., Souček M. (2008): The Research and Development Information System of the Czech Republic. In: *Get the Good CRIS Going: Ensuring Quality of Service for the User in the ERA*, Aleš Bošnjak, Maxmilian Stempfhuber (eds.), CRIS2008, proceedings from the 9th International Conference on Current Research Information Systems, Maribor, Slovenia, June 5th–7th, 2008, IZUM, ISBN 978-961-6133-38-8, pp. 125–130.

Contact Information

Tomáš Chudlarský – tomas.chudarsky@ff.cuni.cz

Jan Dvořák – jan.dvorak@ff.cuni.cz

Institute of Information Studies and Librarianship

Faculty of Arts

Charles University in Prague

U Kříže 8

CZ-15800 Praha 5

Czech Republic