

# The Research and Development Information System of the Czech Republic

*Jan Dvořák, Martin Souček*

InfoScience Praha, s.r.o.

## Summary

The Research and Development Council of the Czech Republic runs a national CRIS that contains data on publicly subsidized R&D. It covers the entire research process - from strategic planning, schemes and calls, to projects and institutional research plans, to results. It also contains links between those stages. Data entry is mandated by law. The data is used as a basis for making a budget and for the R&D efficiency evaluation that has taken place every year since 2005.

## 1 Introduction

The Research and Development Information System of the Czech Republic (in Czech: “Informační systém výzkumu a vývoje” – abbrev. “IS VaV”) is a nationwide CRIS with a moderately long tradition. In the Czech Republic, the State plays an important role in research funding. The Czech CRIS documents the whole process of research, from strategic funding planning to the results. It is run by the Research and Development Council. In this article we will give a brief description of the system in its present state, outline its history, and look at its compatibility with CERIF (Jörg et al. 2007).

## 2 Description of the Czech R&D IS

### 2.1 Legislation

The existence and basic responsibilities of the Czech CRIS are anchored in Act No. 130/2002 Coll., on the Support of Research and Development from Public Funds. This is further elaborated in Government regulations. In particular, it is mandated that:

1. Any research that is at least partially supported by the State Budget of the CR is recorded in the Czech CRIS.
2. Before a state funding provider makes support available to research institutions, a batch of fresh data from that provider has to be successfully submitted to the Czech CRIS.
3. Any research marked as successfully completed has to submit results – publications, patents, technologies, etc.

Furthermore, the R&D Efficiency Evaluation, which has taken place on an annual basis since 2005, provides direct motivation for research institutions to submit results.

## 2.2 Uses

The Czech CRIS has the following basic purposes:

1. To support the R&D Council in preparing and negotiating the state budget proposal for R&D.
2. To disseminate information on:
  - Calls,
  - Current research,
  - Past research along with its results.
3. To provide input data for the R&D Efficiency Evaluation.

## 2.3 Scope

The Czech CRIS tracks the following objects:

1. **State budget proposals for the R&D agenda.** Data items: funding provider, financial amount, classification, justification.
2. **R&D support schemes (funding programmes).** Data items: id code, title (Czech and English), goals (Czech and English), start year, end year, government approval data, EC notification data.
3. **Calls for project proposals – (the “VES”).** Data items upon announcement: id code, programme, funding provider, dates (call open, call close, selection close, selection publish), description of expected proposals and selection criteria (Czech and English), contact information, proposal submission information, financial amounts planned (total and broken up into years). Upon closing: Status, numbers of proposals (received, evaluated, accepted), financial amounts granted (total and broken up into years), failure reason.
4. **R&D projects – (the “CEP”).** Data items for in-progress projects: id code, title and short description (Czech and English), programme, funding provider, classification (primary up to tertiary discipline, keywords), start date, end date, status (beginning, running, ending, one-year, stalled, completed, stopped), participating organizations (including individuals – for small projects) with roles, participating individuals with roles, financial amounts (overall and state funding; planned per project and per participating organization; total and broken up into years). For completed or stopped projects, one tracks the data items listed above plus actually spent funding per project and per participating organization, plus project evaluation (excellent, successful, unsuccessful; Czech and English evaluation by the funding provider), minus participating persons.
5. **Institutional research plans (a form of institutional R&D support) – (the “CEZ”).** Similar to R&D projects, but broader and larger in scale.
6. **The results – (the “RIV”).** Data items common to all types: result type, language, title and abstract (Czech, English and the language of the result), submitting organization, authors (names, id, nationality, flag for affiliation with the submitting organization), discipline, keywords. Data items for publications (monographs, articles, contributions in conference proceedings): ISBN, ISSN, journal or book or proceedings title, edition, page count, page range, year of publishing. Data items for patents: patent number, patent office, patent owner, dates (registration and issue). Data items for new technologies: identification, owner, technical and economic parameters.

Indirectly it also collects information on:

7. The organizations that are active in R&D – universities are tracked to the faculty level.
8. The researchers – as project coordinators, as project team members, and as authors.

## 2.4 Architecture

### 2.4.1 Data Collection

The state funding providers are the direct partners for data submission. Data input into the Czech CRIS is strictly batch-oriented. The newest version of a data batch is considered valid, so newer batches overwrite the older ones. This keeps the responsibility for the data contents of the CRIS indivisibly within the hands of the state funding providers.

Before it can be submitted, any data batch must pass a set of formal checks. These checks are available in the form of a webservice; anyone may use it to check their data. The data format is XML and there is an XML Schema defined for each year's data structures. In addition to the XML Schema, there are nearly two thousand data integrity checks the data has to pass.

Data about projects and institutional research plans are collected on an annual basis, forming a series of snapshots that track the potentially multi-year processes. Once the project or the institutional research plan is completed, another snapshot is collected that summarizes the whole process and outcome.

The definition of the collected data set is renewed each year. The CRIS thus encompasses new data items to be collected, refinements in pre-existing data items or classification schemes, and the (rare) drops of data items that are not to be collected any more.

The larger-scale state funding providers produce XML data batches about projects as exports from their internal information systems. Smaller-scale providers are given a tool to edit the data interactively (the "Vklap"). This tool integrates all the data integrity checks that the data is required to pass and provides the often useful functionality of merging data files. Using this tool, one can support complex data collection arrangements of up to three intermediate stages.

### 2.4.2 Data Processing

State funding providers submit the data batches that pass the formal checks to the R&D Council. Here the data is checked independently and recorded in the Central Database. All current data batches are processed at regular times. Processing has the following stages:

**Data consolidation.** The data was collected in more than 60 different formats. The first task is to convert all of this data into a single data structure that uses a single set of classification schemes. The input data was in DBF files until 2005, and since 2006 only XML has been used. Data that accumulated in previous releases of the Central Database (up to 2001) is transformed differently, as one large batch.

**Data integration.** The common data structure consists of 18 database tables. The data then needs to be cleansed, normalized, and integrated. Cleansing of the consolidated data makes use of external data from the Czech Statistics Office (the organizations register) and of other, internally maintained auxiliary data. For multi-year entities (projects and institutional research plans), one has to construct a current state-of-the-art view. Older data has gaps that

need to be filled in systematic ways in order to make the data consistent. The fully integrated data structure consists of 25 database tables.

### 2.4.3 Data Presentation and Dissemination

The following types of output are produced from the central database:

**Work tables for the R&D Council.** There are around 20 work tables that are used to produce both repeating and ad-hoc reports. At a defined time some of these tables are used as the input for the R&D Efficiency Evaluation.

**Web presentation support data.** This is the underlying data for the web presentation application running at <http://aplikace.isvav.cvut.cz/>. This application lets users search the data according to various criteria.

**Support data for the data collection process.** In some cases when a new data batch is to be based on one from the previous year, the state funding providers are given a data skeleton that just needs to be filled in. This saves the funding providers without an internal information system a lot of copying and pasting.

## 2.5 Technology and sizes

The Czech CRIS uses the Oracle 10g database and Java SE and EE technologies. In addition to vanilla SQL, the database makes use of the XML processing capabilities of the Oracle database as well as analytical functions. The web presentation application is implemented as a Java EE application. It runs in a JBoss application server. The data entry tool and the central database management application are implemented as Java SE applications with Swing graphic interfaces. They are deployed using Java Web Start technology that takes care of updating the application code and resources to the current version. The data entry tool is supported by several megabytes of read-only data; the Hsqldb database is used to make this data available through SQL.

Data migration and other ad-hoc data manipulations are carried out using the Ant build environment which has been enhanced through several custom tasks for data manipulations. DBF files are used as the intermediate data storage format. The data outputs for the R&D Council take the form of around twenty DBF tables that are packed in ZIP archives. This packing takes place along with the file production – this streamlining effect is achieved with the open-source Datasink tool (Datasink).

The database model is maintained using the Toad Data Modeler database design tool (formerly known as CaseStudio2). It is now comprised of over 500 entities, of which around 300 are in active use.

The database size is currently around 40 GB with an increase of ~4 GB/year. The table below shows the numbers for some basic objects in the database.

*Table 1:* Numbers of main objects in the Czech CRIS (April 2008)

| Data object   | Current number | Increase per year | Collected since |
|---|----------------|-------------------|-----------------|
| R&D projects  | 30,000         | 1,800             | 1994            |
| R&D project phases (year-based snapshots)                 | 95,000         | 7,500             |                 |
| Institutional research plans                              | 880            | 0                 | 1998            |
| Institutional research plan phases (year-based snapshots) | 4,700          | 300               |                 |
| R&D result records (publication, patent, technology)      | 530,000        | 65,000            | 1998            |
| Cleansed R&D results                                      | 420,000        | 50,000            |                 |
| Calls   | 400            | 30                | 2000            |
| Research organizations                                    | 4,200          | 230               | –               |
| Project coordinators                                      | 26,000         | 900               | –               |
| Schemes   | 140            | 6                 | –               |

Full data processing takes between 2 and 4 hours on the production machine (a six-CPU partition of an IBM P-560 with 48 GB of RAM running over a fiber-channel disk array of SCSI disks, non-dedicated). Queries from the web application are usually processed well into the sub-second range, with the exception of results, where queries take around 5 seconds.

### 3 History

The Czech CRIS started with the launch of the grant system in 1993. In the beginning it was a single data table of current projects. With the first projects finishing, the problem of recording past research had to be solved. In 1995, the information system took the form of a FoxPro application. In 1997 the system went through a major rewrite, with Informix 7 as the database engine. Since 1998, the project data and the institutional research plan data have expanded into three interrelated tables, one master and two details, and data about results and its connection to research projects has started to be collected, too. In 2000, data about calls was added.

An effort to integrate data between the information system components started in 2001, which resulted in another major rewrite. XML for the input data was introduced in 2002, and fully adopted in 2005.

In 2004, the decision was taken to move the information system from the Office of the Government to the Computing and Information Centre of the Czech Technical University. The years 2005-6 brought about yet another major rewrite, this time also switching the database technology to Oracle.

We hope that the need for the next major rewrite doesn't come before 2011.

## 4 The Czech CRIS and CERIF

Part of the motivation in most of the rewrites of the Czech CRIS was to make the database structure closer to the CERIF standard. The present state is that a CERIF export from the Czech CRIS is feasible and can be implemented with moderate effort.

However, the full set of CERIF entities is not completely covered in the Czech CRIS. It does not record researchers' CVs, qualification, expertise and skills, research interests, prizes, facilities, equipment, and services. It only records past events as a subclass of results.

When trying to match the semantic layer of CERIF, the following incompatibilities would need to be overcome:

**Currency.** The Czech CRIS, from its very inception, expresses financial amounts as multiples of 1,000 CZK. This is not a constant value in time: it was ~ 30 EUR in 2002, and ~ 40 EUR in 2008.

**Discipline classification scheme.** Since 1998 the Czech CRIS has used a custom discipline classification scheme consisting of 10 main areas and 123 disciplines.

**Language codes.** The Czech CRIS uses the ISO 639-2 scheme (the three-character codes).

**Country codes.** The Czech CRIS uses the ISO 3166-1 alpha-2 scheme (the two-character codes).

## 5 References

The R&D Council of the Czech Republic: The Czech Research and Development Council Website. The English version. <http://www.vyzkum.cz/?lang=en>.

The data of the R&D Information System of the Czech Republic. The English version. [http://aplikace.isvav.cvut.cz/locale/en\\_US/](http://aplikace.isvav.cvut.cz/locale/en_US/).

Jörg, B.; Jeffery, K.; Asserson, A.; van Grootel, G.; Grabczewski, E. (2007): CERIF2006-1.1 Full Data Model (FDM) / Model Introduction and Specification. <http://www.eurocris.org>.

Datasink. An Open-Source Software project on the SourceForge.net site. <http://sourceforge.net/projects/datasink>.

Dvořák, J.; Souček, J.: The Research and Development Efficiency Evaluation in the Czech Republic. Companion article, CRIS2008 conference.

## Contact Information

*Jan Dvořák | Martin Souček*

InfoScience Praha, s.r.o.

Dlážděná 4

CZ-11000 Praha 1

Czech Republic

e-mail: [jan.dvorak@infoscience.cz](mailto:jan.dvorak@infoscience.cz) | [martin.soucek@infoscience.cz](mailto:martin.soucek@infoscience.cz)