Title: CroRIS Data Quality Issues and Possible Solutions

Presenters: Petra Udovičić *, Ognjen Orel *
* University Computing Centre (Srce), University of Zagreb, Croatia

Kind of activity: Presentation

Abstract:

Previously, we have reported on the progress being made in the development of Croatian national CRIS (CroRIS), its architecture, design, modules and the role it will take as a NOSC service catalogue as well. Little has been told about the data quality, especially taking into account the various sources of data being migrated and blended into the system and a large number of everyday users.

During the initial phase of system development, several challenges emerged, that required the focus of a significant part of the development team. One of these was the data quality of several important data sources that needed to be completely migrated to CroRIS. Migrating the first one, the official records of the Ministry of Science and Education (MSE) regarding researchers and institutions mostly meant the adaptation of the data to the data model that was used in CroRIS, a CERIF hybrid. Migrating the others meant not only the adaptation to the data model but also to the data that was already in CroRIS. The biggest problems were, expectedly, the identification of key entities, such as persons, and the reconciliation of the semantics. Each of these systems, being built over the years, had its design problems which only helped grow the data entropy. A large part of the data was reconciled successfully, but there are some records which couldn't be identified properly and will pose the incorrect data.

Since CroRIS is designed to promote openness, most of the data is available publicly. Personal data, including the official national records of MSE, and all of the connected entities (projects, publications, citations, etc.) are available to those persons. Therefore, the awareness of data quality is disseminated to all, and all have permission to edit and add new data related to them.

But will this help improve the data quality? Given the fact that CroRIS is a single-installation live system with around 30.000 users, one can remain sceptic about the future of data quality. To tackle this issue, several measures were introduced in the system: role separation, scope limitation and data verification. Role separation and scope limitation are orthogonal, creating the system permission matrix.

A complex set of roles is in place. The basic role is given to anyone who can log in to the system. Those are the people who have their academic electronic identity in the national academic AAI system, having the affiliation of an employee (researcher, administrator) or PhD student. Users with the basic role can change their personal data (but not all, if they are already part of the official registry of the MSE) and add or edit the publications they authored, projects, equipment, services, etc. they are affiliated with.

In each of the institutions (research institutes and higher education institutions) taking part in CroRIS (around 200 of them), there are several editors. These editors are the users who are trained and specialized to maintain some of the data in the institution, e.g. the publications data (mostly librarians), projects data (mostly the project office), personnel data (mostly HR), etc. This is where data verification comes in. As each user can enter or change the data they are entitled to, the data still needs to be verified. There is a small portion of data that does not require verification (such as personal contact

data), but mostly all the data needs to be verified. The verification is done by the editors in the institution. The publications' editors check, correct and verify the data about publications, the projects' editors do the same for the projects and funding data, etc.

The editors are appointed by a CroRIS coordinator in each institution. This is a person who manages CroRIS operations at the institutional level (mostly on a vice-dean, or assistant director level). They can autonomously give the editor roles for different scopes of data in the institution, including themselves. The CroRIS coordinator is formally appointed by the dean or director of the institution.

There are also admins, editors specialized for specific parts of the system, but on the system level. They are bound to the system operations team and serve as editors for those institutions that currently don't have appointed editors, or as mediators when necessary. Like the projects' editor, a projects' admin will verify and correct the data about the projects, especially if the project is bound to multiple institutions.

The super-admin role is the coordinator of the whole system, those are the member of the development and maintenance team. Super-admins maintain the semantic level of the data and other top-level roles, such as admins or CroRIS coordinators.

On top of these roles, there are some other specific roles in the system, like the MSE employee roles. Those are the users with special permissions regarding the official registers of MSE, who moderate and verify the specifics of the researchers and institutions. Another set of specific roles is envisioned in the future, mainly regarding the reporting subsystem. Those could be the members of national bodies, councils, university managers and alike.

In this talk, we will give a detailed view of the CroRIS role and scope limitation system. The special focus will be put on the data verification process and some results of the process will be shown. The future work on CroRIS and the data quality endeavours will also be addressed.

CroRIS is being created by the University Computing Centre, University of Zagreb (Srce) as a part of the Scientific and Technological Foresight project (STF), led by the Croatian Ministry of Science and Education, with the help of the Centre for scientific information of the Ruđer Bošković Institute. It is a CERIF-based national research information system that integrates a large amount of data regarding research in Croatia, including data about researchers, institutions, projects, publications, products, patents, equipment, services etc.