Ongoing Development of the CERIF[1] Research Information Metadata Model

CONTACT INFORMATION

Jan Dovrak (InfoScience Prague), CERIF TG project leader: jan.dvorak@infoscience.cz

Ed Simons (Radboud University, NL), President of euroCRIS (e.simons@ru.nl).

The project is executed by a Task Group (CERIF TG) of "euroCRIS": an international network of research information experts, developers and users (www.eurocris.org).

OVERVIEW AND GOALS

The overall aim (and challenge!) is to develop a comprehensive metadata model, covering as much as possible the full domain (all aspects) of research information. In a nutshell:

- *Persons* involved in the research and their possible roles (researcher, manager,…)
- *Organizations* involved: research institutions, funding agencies, partners, customers,…
- *Project descriptions*: content description, academic domain, period, language,…
- *Input:* both in terms of human (fte's) and financial resources.
- *Output*: publications, patents, products, contributions to events, services rendered,…
- *Datasets used/produced:* identification*,* volume, location, access/use rights,…
- *Equipment, tools and/or services used* in the research.
- *Metrics*: number of publications, citations, grants, awards, etc…
- *Impact* on society: politics, culture, wealth creation, well-being of people…
- *Semantic definitions:* controlled vocabularies, both for formal-administrative (e.g. typologies of roles in research or publication types) and content-related (e.g. taxonomies for a given scientific domain) characterization and clarification.

Core principles for the development of the model are:

- *As complete as possible* (see above).
- *Optimal application of the relational model*: high nominalization and expression of aspects of entities as relations (linking tables) instead of attributes (in the entity table).
- *Optimal (fine) granularity:* the model should be able to produce (any other existing metadata format in the

[1] CERIF, the: "Current European Research Information Format", is an official EU recommendation to Member States. euroCRIS has been mandated by the EU for the (further) development and maintenance of the model.

- research information domain (DC, CKAN, MODS-DIDL,
  OpenAire, VIVO-ontology, etc…) and thus be interoperable with any of these formats.
- *Multilinguality and Periodization* of all entities and relations (language as well as start- and end-date field).
- *Separation of semantics and data:*(accommodation of semantics in a separate "Semantic Layer"*,* allowing the expression (linking) of multiple semantical aspects for any given entity or relation, as well as the mapping of various semantical schemas (e.g. various taxonomies).

STRUCTURE OF PROJECT

The project is a continuous ongoing core activity of euroCRIS. As said it is executed by a "CERIF Task Group", in which IT experts and domain experts from various countries (mainly Europe), and representing both the user and developer (vendor) communities, work together to advance the model.

MAIN FINDINGS & SUCCESSES

Various countries in (Western) Europe have implemented (sometimes already for a long time, e.g. the Netherlands, Denmark) or are in the process of implementing (e.g. UK, Germany) research assessment processes and formats to evaluate the return on their taxpayers money invested in research. Typically these assessment formats measure the relation between input and output of research, based on specified data supplied by the research institutions. In order to be able to comply efficiently to this assessment demand, in most of these countries so-called CRIS systems (current research information systems) have been implemented by universities and research institutes. Nowadays the CERIF datamodel is the accepted international standard in Europe for the development of CRIS systems and as such is the main source for the measurement/assessment of the scientific performance of research groups. Currently several hundred of CERIF-based CRIS systems have been implemented in Europe.

Another very concrete contribution of CERIF to the measurement of science is embodied in a project that euroCRIS currently is executing in cooperation with Elsevier to "CERIF-y" the so-called "Snowball Metrics": a set of measurements for scientific performance developed by Elsevier in cooperation with universities in the UK.

Apart from these practical implementations of CERIF one can say that the model, given its broad scope mentioned above, in general holds a lot of the metadata information needed for measuring science and scientific performance. Important is that CERIF holds a lot of *contextual metadata* of research allowing to put measurements and evaluations of science based on other resources (e.g. publication and citation databases) in the correct perspective (e.g. to really judge the value of a number of publications by an institute, one also need information on the number of researchers or available resources in general of that group).

MAIN CHALLENGES & NEXT STEPS

The CERIF metadata model is already fairly complete. Main things to be added are:

- Additional metadata for datasets: in this respect euroCRIS has a leading role in the metadata working group of the Research Data Association (RDA)

- Fine-tuning of the metadata for measuring (societal) impact and specifically the datavalues for these metadata.

Main challenges within the broader framework of the research information community:

- Completing the agreed upon joint metadataset by euroCRIS and the OpenAire community for interoperability between OA Repositories and CRIS systems.
- Completing the mapping of CERIF to the VIVO-ontology. (both projects are already well under way, but must be finalized)
- The development of a reference "pure CERIF" (system) implementation on the euroCRIS web site, in order for vendors and other developers to test their CERIF-compatibility.
- Starting up projects with "the business", i.e. the communities of various academic/research domains to define the content-related semantics for research metadata (controlled vocabularies) . This is a major and huge challenge, but in the long term probably indispensable for optimal discovery of information in the "big data" era.