Information systems of research funding agencies in the "era of the Big Data". The case study of the Research Information System of the European Research Council

Alexis-Michel Mugabushaka, Theodore Papazoglou
European Research Council Executive Agency, Brussels[1]

Summary

The European Research Council (ERC) was established in 2006 to support and strengthen excellent frontier research in Europe. Since its beginning, it has received and evaluated over 32,000 proposals and funded more than 2,500 projects (as of June 2012). To assist in its effort to account for the added-value of its funding activities, the ERC is developing a CERIF-compliant-research information system. It is conceived as an inte-grated platform which combines several tools to support the gathering, management and analysis of data on results of funded projects. The distinctive features of the ERC system is that it makes extensive use of semantic web opportunities and machine learning techniques to minimize the burden on funded researchers. In this paper, the rationale and main functionalities of the system are presented.

1 Introduction and Background: The European research council

The ERC is a European research funding body set up to support frontier research in Europe through competitive, investigator-driven grants. ERC implements the "Ideas" Specific Programme of the EU's 7th Research Framework Programme and have an overall budget of € 7.5 billion over 7 years (2007-2013). It consists of an independent Scientific Council (ScC) which is responsible for the development of the scientific strategy and decision-making and an Executive Agency (EA), which deals with all aspects of administrative operations and programme execution.

Within the 7th Research Framework Programme, whose overriding aim is to "*contribute to the Union becoming the world's leading research area*" (EC 2006a, p.4), the objective of the "Ideas Specific Programme" – and thus the mission of the ERC – is to *"to reinforce excellence, dynamism and creativity in European research and improve the attractiveness of Europe for the best researchers*". It follows an *"investigator-driven approach, (supporting) 'frontiers research' projects carried out by researchers on subjects of their choice*" (EC 2006b, Annex I).

As of June 2012, the ERC Scientific Council had developed five funding schemes to translate the ERC mission into reality (ERC WorkProgramme 2013).

[1] The views expressed in this paper are the authors'. They do not necessarily reflect the views or official positions of the European Commission, the European Research Council Executive Agency or the ERC Scientific Council.

- *ERC Starting Grant* supports researchers who are establishing their own independent research team or programme.
- *ERC Consolidator Grant* supports researchers at the stage at which they are consolidating their own independent research team or programme.
- *ERC Advanced Grant* supports projects directed by leading advanced investigators of whatever age, who are already well-established in their respective research communities.
- *ERC Synergy Grant* support teams of two to four individual Principle Investigators (PIs) whose project depends on and profits from their complementary expertise.
- *ERC Proof of Concept* supports ERC grant holders who wish to identify a development path and/or an IPR strategy for ideas arising from their ERC-funded projects.

Since the starting of its activities, the ERC has launched 12 grant competitions yielding a total of over 32,000 proposals. About 2,500 projects have been selected (in the 8 calls which are completed as of June 2012).

The establishment of the ERC has raised high expectations in the scientific community and in the research policy circles. This makes it necessary to design an evaluation and monitoring approach to assess if those expectations are met and to inform discussions about corrective measures.

This paper reports on the efforts of the ERC in creating a research information system: the Research Information System of the ERC. It is conceived as an integrated platform which combines several tools to support the gathering, management and analysis of data on results of funded projects. Its distinctive feature is that it aims to take advantage of the opportunities presented by the huge amount of data available online and to use advanced machine learning techniques in the analysis of those data in order to minimize the burden on funded researchers.

The paper is structured as follow. After this section which briefly presented the European Research Council, section 2 puts the research information system in wider context and discuss the experiences of research information systems of other funding agencies. Section 3 presents the key features of the ERC system and its main functionalities. The final section discusses challenges and present planned next steps.

2 Changing expectations on Research Information Systems of Research funding agencies

2.1 The evolving concept of "accountability"

The recent debates on how to overcome the on-going economic crisis have revived and reaffirmed the consensus on the economic value of Research and Development (R&D). In the majority of advanced and emerging countries, national and regional governments have devised ambitious strategies to stimulate growth by R&D and are making significant investments in R&D systems. This policy is based on empirically strong evidence showing that long term economic health of countries rests on their contribution to global scientific and technological advances.

Research funding agencies play an important role in this context. In a simplified description of their place in national and international innovation systems, they operate under the principal-agent model whereby public authorities entrust them with the public funds (i.e. taxpayers' money) and mandate them to autonomously allocate those funds to most promising research undertakings.

As "agents", the research funding agencies have always faced a demand to account for the public money they manage and they have developed different approaches and tools to assist them meet this demand. A 2009 report by the European Science Foundation (ESF 2009) describes the approaches taken by national funding agencies in Europe. The report shows that in ex-post evaluation of funding schemes and research programmes, national funding agencies in Europe have a lot in common, despite some differences due to the various missions they have.

For the technical infrastructures we can see that, since the beginning of 1980s, research funding agencies started creating databases of research projects. The idea of a Common European Research Information Format (CERIF), was conceived at the same time with the ambition to enable the exchange of information and inter-operability of those systems (EuroCRIS 2012).

A glance at the development of the CERIF Format (Jörg 2008, p. 184) illustrates well how the data model evolved over the years to accommodate the increasing and changing nature of the information that research information were expected to manage.

The first release of CERIF, published in 1991, included only one core entity: "Research Projects". The CERIF 2000 release added person and organisations to the core entities and publications and patents as secondary (2nd Level) entities. The CERIF 2006 reclassified publications as core entities. The CERIF 2008 and the subsequent releases (CERIF 1.3 and CERIF 1.4) substantially revised previous versions adding new entities added and regrouping all CERIF entities in logically coherent categories. Schematically, the new CERIF Release has following entities:

- Base Entities : Project, Person, Organisation
- Results entities : Publication, Patents and Products
- 2nd Level entities: which are important information points related to base entities and results entities such as Prizes (relating to entity People), Citations (related to entity Publications), Information on funding scemes (related to Projects)
- Infrastructure entities: which records research facilities, equipments and services.
- Link entities : relationships or links between CERIF entities)

This expansion of informational entities included in the CERIF model and its increasing complexity is a good indicator of the changing expectations on research information systems of Research funding agencies. It illustrates how the concept of "accountability" has shifted and its dimensions expanded. While its core question remains the same ***("what have you done with the money"?***), the expectations on the reporting have evolved away from simplified model ***("which projects do you fund ?")*** to a more sophisticated model in which emphasis is put on research results and ultimate research outcomes ***("which results have you achieved?"***) and "***what is their significance – for the science, the economy and the society ?"***).

This reflects a broader trend to justify public investments and monitor their returns which permeates the debates about the economic and societal impact of Research and Development investments (R&D). Policy makers and the public are asking not only **if** R&D investments produce benefits but also **how**. This makes it imperative for research funding agencies to identify and strengthen the mechanisms of resources allocation which promises to maximize the investment benefits. Their research information systems are expected to include not only funded projects but also their results. They are expected to explain their merits and to inform the public – which ultimately funds this research – on the achievements of those efforts in an accessible format.

With respect to the scope of research information systems of funding agencies, we can distinguish – at a high abstraction level - three dimensions of "accountability":

- **Reporting**: this refers to the initial focus of research information systems. To act as project databases and provide to outside users basic information on the funded projects (such as topics and summary of projects as well as people and organisations involved).
- **Programme Evaluation**: this refers to the need to include information that allows a systematic and objective assessment of the achievements of funding agencies and whether the stated objectives of their funding schemes are met. This includes project results (publications, patents, etc …) as well as quantitative and qualitative information which allow to put them in context (benchmark, understand the significance of discoveries made.)
- **Public information**: this dimension complements the two previous ones. The target group of information provided in the first two dimensions is likely to be policy makers and other interested parties who can use this information. The public information dimension takes into account the fact that the "public" – who ultimately fund and benefits from the research supported – are important stakeholders in the accountability process. There is a need therefore to provide information on results of the public investments in research in a format and language easily understandable by the general public. This includes for example short publications ('snippets') highlighting scientific discoveries and their relevance.

Figure 1: The three dimensions of 'accountability'

2.2 Potential of the "era of Big Data"

The conceptual framework of the accountability concept provided above allows a differentiation of the data in the research information systems of research funding agencies.

The data needed for the "Reporting dimension" are generally taken from operational databases in which results of the selection process as well as financial transactions are recoded. An analysis of

the information systems of European research organisations conducted by the European Science Foundation in 2009, show that most research information of European funding agencies included only information related to this dimension (ESF, 2008, p. 7). For the two other dimensions, the data have to come to external sources. Traditionally, those data are recorded in progress-report that funded researchers are requested to submit to their funders periodically. Until recently, most agencies requested researchers to file those in an unstructured way (either on paper or in PDFs document) and the rich information they contain are in such cases not easily usable.

Currently, funding agencies have started to collect this information in a more structured way, generally by offering a web-based interface in which researchers enter those information. Some funding agencies use this approach to collect results of the research continuously or periodically and update it two years after the grant termination (ESF 2011).

This offers an advantage over the traditional gathering of information at mid-term and at the end of the project. It allows the organisations to gather data which are easy to process and enable them to report with updated data on a regular basis. However, the increasing demand of data needed in increasingly great detail will inevitably add to the burden put on principal investigators. It creates a tension between on one hand, the legitimate need of the funding agency for information to better account for public funds and the researchers, who are more and more requested to spend a great share of their time in administrative task related to the grant they received. A now famous example of the burden put on researchers is the finding of a survey among 6,081 faculty members at selected US research institutions. The authors estimate that, after winning a research grant from federal funding agencies, they spend slightly over 40% of their time, not on active research on the grant, but on post-award administrative activities such as grant progress-report submissions. (Decker , 2007, p. 17). The task they view as most burdensome is the "grant progress report submissions" (ib. p. 19).

The recent ICT advances – especially in what has been coined the "era of big data" - hold great promises to overcome this tension. The concept "big data" refers to the huge and increasing amount of structured and unstructured data which are available online or in internal databases of organizations as well as the technologies which have been developed to access and process this information.

McKinsey popularizes the term "era of big data" in highlighting how companies use this kind of data to maximize their return on investments (McKinsey 2011). They predict that this practice will be a "game changer" for virtually all sectors and that arising opportunities will create new industries.

Research information systems are in integral part of this "big-data" trend and have witness a genuine revolution in last years. To name just a few emerging technologies:

- The digitalization of scholarly communication (and the related open access movement) has allowed millions of publications to be made accessible online.
- Research organisations publish profiles of their researchers online and routinely publish highlights of their achievements (e.g. press releases on important discoveries) and the "interactive internet" allow researchers to share information, create new communication channels (blogs, comments …).
- Several commercial and non-commercial parties offer databases on publications and patents. There are promising efforts to create unique "author identifications" (e.g. the

ORCID[2]) and unique identifications of publications such (digital object identifiers) which will help the users of such systems.

- The popular science communication has also changed: several "news aggregator" are competing with popular science magazine to offer content in a user-friendly way
- Powerful tools have developed to make it easier to use this information. The "semantic web" technologies allow users to access the data online. "Data Mining" helps extracting information from seemingly unstructured data, and " Machine learning" allows the use of small sets of "trained data" - for example processed manually by human experts - and infer the patterns and regularities to a larger body of data which cannot be processed manually.

Several research funding agencies are starting to take advantages of those advances to meet their accountability obligations and minimize the burden put on funded researchers.

- The Research Portfolio Online Reporting Tools (RePORT)[3] of the National Institutes of Heath provides information on funded projects and their results. It captures automatically publications (from PubMed) and use text mining to cluster research projects.
- Arguably, by far the most ambitious initiative is the STAR Metrics. This is an effort by several US federal research funding agencies to develop an information infrastructure which will help to document the return on public R&D investment. STAR Metrics is designed as a two-phase project. The First level aims to gather information on workforce in funded projects, using administrative records of universities and retrieving the information in an automated way. A pilot project of level 1 was successfully accomplished for six universities in 2009 (Lane, 2010). The second level will collect research outputs as well as other information relevant to understand the impact of funded research (taking into account the diversity of traditions and practices in research fields).

The ERC research information system, although different in design and in architecture, can be seen as part of this trend. Its ambition is to take advantages of the best amount of data already in public domain or from commercial sources to assist it accounting for the added-value of its funding activities while at the same time minimizing the burden on funded researchers.

3 The ERC Research Information System: Key features and functionalities

The research information system of the ERC is designed to support the monitoring and evaluation strategy of the ERC. It is clear that no research information system alone, however ambitious it might be, can document the funding activities and all their results and account for its impact in all its facets. It will take years of analysis and long term studies to understand and document how the impact of ERC funding unfolds. A well designed research information systems can help those studies. It can prevent the duplication of data collection exercises (and thus minimizing costs of evaluation studies and burden on ERC funded researchers). It can offer good quality data which are a crucial foundation of any successful monitoring and evaluation strategy.

[2] http://about.orcid.org/

[3] http://projectreporter.nih.gov/reporter.cfm

3.1 Data Model

Although the long term goal of the ERC Research Information System is to enable the ERC to collect systematically data which can help in analysing the impact of the ERC in all its facets, in its initial phase, it will concentrate of four types of information. The chart below, gives a schematic representation of the data model.

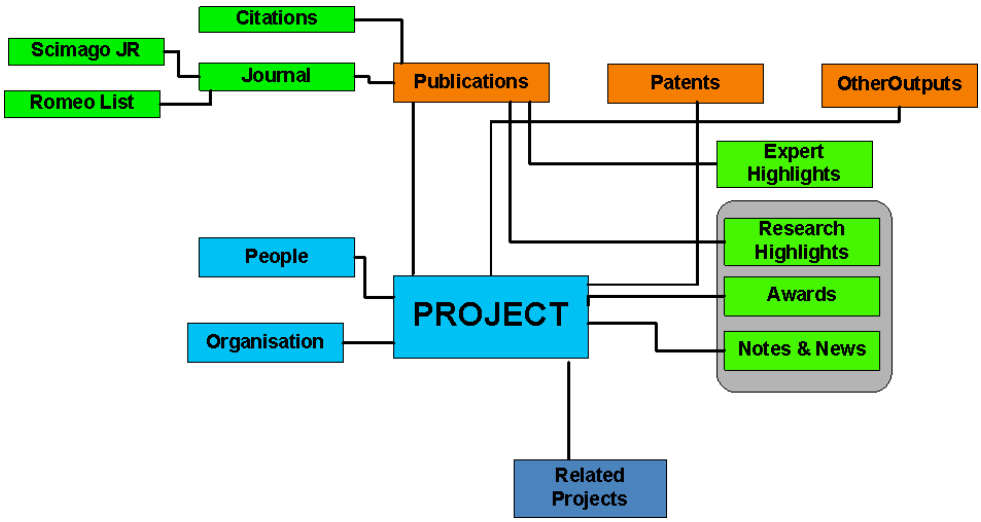

Figure 2: Schematic representation of the data model of the ERC Research Information System

1. **Information on projects**. This includes the entity "project" with information like title, abstracts, research areas, starting and ending dates; entity "People" for information researchers involved in the project and the entity "organisation" on involved research organisations.
2. **Results of projects**. This will include publications (in first phase only journal articles, conference proceedings and monographs) as well as patent applications.
3. **Quality and Impact of Results**. In addition to the traditional measures used in bibliometrics (article-level metrics such as citations and journal level-metrics such as relative rank of journals), this will include more qualitative information on the advances those research results make.
4. **Information for the general public**. While the information categories listed above are likely to be of interest for many stakeholders, they may not inform the general public on the relative importance of discoveries made in the projects.

3.2 Data Sources

As already mentioned, the ERC Research Information System aims to minimize the burden put on PI in collecting the data by making use of already existing data sources. In the initial phase following data sources will be used.

Information on projects

The data will be retrieved from ERC operational database. In addition, it will include also information on clusters of projects. This will be achieved by applying machine learning algorithms to identify thematically related projects

Results of projects

For journal articles and conference proceedings, the data will be collected from free online services (PubMed and ArXiv) and from commercial databases (Thomson Reuters Web of Science and Elsevier's Scopus). Data on monographs will be gathered from the "Library of Congress" and the German National Library. Provision will be made to collect "other type of results" in flexible manner. Text-mining will help data quality efforts (e.g. attribution of article to author or linking it to a project)

Quality and Impact of Results

The system will make use of post-peer acceptance peer-review which is done in some journals such as Nature journals ("News&Views") or Science ("Perspectives") and it will also track research results which are highlighted by authoritative sources ("editors' choice in Science), publications winning awards from professional research societies etc ..). It is planned, in future release, to consider systems such as "Faculty of 1,000".

Information for the general public

The systems will make use of the popular science press (and related online news services) and the press releases of research institutions to present highlights of the results in format easily accessible to the public.

3.3 Architecture and Implementation

Architecturally, the ERC Research Information System is designed as a platform of several modules. The database modules consists of a staging database, in which data preparation steps such as data cleaning and validation take place and a production database in which the data are finally stored and which is used for the reports and analysis.

The production database module (and related web-based interface) is built on the basis of CONVERIS, a research information system[4]. The specificities of the ERC system required however a relatively long customization phase. The production database store the data on the informational entities listed in section 3.1 offer reporting capabilities. In addition, it includes processes to download automatically data on publications, citations and journals.

All other functionalities of the system are included in the staging system, which is developed with ERC internal resources.

[4] The system has been selected following a public call for tenders. For information on Converis, see: http://www.avedas.com/en/converis.html

The first prototype of the system has been rolled out in April 2012. In the initial phase the system will remain for internal use only but future releases might include functionalities to interact with ERC Grantees and to make selected datasets available online.

4 Concluding remarks and Outlook

This paper describes the rationale, data model and implementation of the ERC research information system. It has been designed to support ERC monitoring and Evaluation Strategy and aims to take advantages of the opportunities to "the era of big data" to collect data automatically from available sources and thus minimizes the burden put on ERC Grantees.

The first prototype of the system has been rolled out in April 2012. In the initial phase the system will remain for internal use only. ERC will carefully evaluate the system and discuss with stakeholders areas which need to be optimized.

One challenge is to keep up with the rapidly changing landscape of information systems. We can expect that new data sources will be created and existing ones will disappear or radically change their business models. The data sources will be regularly reviewed and updated and it needs to be flexible enough to keep up with those changes.

In particular, the opportunities presented by institutional repositories should be explored. Indeed, many research organizations have developed sophisticated research information systems which systematically – if not exhaustively – collect scientific records of their researchers. In most cases, they make use of the same data sources as those listed in section 3.2. Future releases of the ERC research information system should seek how to harness this potential in a synergetic manner.

In the long run however, the acceptance of the systems will depend on its ability to deliver what it has promised. In particular, the extent to which it helps ERC reporting its achievements to the public authorities and to the public, it reduces the burden on researchers and, finally, how it will help studying the impact of ERC funding.

References

Bosnjak & Stempfhuber (eds.) (2008). Get the Good CRIS Going: Ensuring Quality of Service for the user in the ERA. 9th International Conference on Current research Information Systems. Maribor: Institute of Information Service.

Brown, B. et al. (2011).Are you ready for the era of ‘big data’? McKinsey Quartely October 2011

Decker, R et al. (2007). A profile of federal-grand administrative burden among federal demonstration partnership faculty. A Report of the Faculty Standing Committee of the Federal Demonstration Partnership, January 2007.

EC (2006a): Decision No 1982/2006/EC of the European Parliament and of the Council of 18 December 2006 concerning the Seventh Framework Programme of the European Community for research, technological development and demonstration activities (2007-2013);

EC (2006b): Council Decision of 19 December 2006 concerning the specific programme: "Ideas" implementing the Seventh Framework Programme of the European Community for research, technological development demonstration activities (2007 to 2013); 2006/972/EC.

ESF (2011). The Capture and analysis of research outputs. Working Document of the Member Forum on Publicly Funded Research.

ESF (2009). Evaluation in National Research Funding Agencies: approaches, experiences and case studies A report of the ESF Member Organisation Forum on Ex-Post Evaluation of Funding Schemes and Research Programmes

ESF(2008): Windows to Science. information Systems of European Research Organisations. Report of the EUROHORCS- ESF Working Group on Joint Research Information System. Starsbourg: European Science Foundation.

EuroCRIS (2010). CERIF Introduction. http://www.eurocris.org (retrieved April 2012).

Jörg, B. (2008), CERIF: Common European Research Format - insight into the CERIF 2008 Release. p. 183-192, In Bosnjak & Stempfhuber (2008).

Lane, J. (2010). Let's make science metrics more scientific, Nature 464, p. 488-489 (25 March 2010)

Contact Information

Dr. Alexis-Michel Mugabushaka

European Research Council Executive Agency
Unit A1 : Support to the Scientific Council
Office: COV 24/161
B-1049 Brussels

alexis-michel.mugabushaka@ec.europa.eu

http://erc.europa.eu

Dr. Theodore Papazoglou

European Research Council Executive Agency
Unit A1 : Support to the Scientific Council
Office: COV 24/165
B-1049 Brussels

theodore.papazoglou@ec.europa.eu

http://erc.europa.eu