CRIS 2014

# Harmonising Research Reporting in the UK – Experiences and Outputs from UKRISS

Brigitte Jörg[a,b,]*, Simon Waddington[c], Richard Jones[d], Stephen Trowell[e]

[a,]*euroCRIS, The Netherlands, [b,]JeiBee, United Kingdom*
[c,] *Centre for e-Research, King's College London, United Kingdom*
[d,] *Cottage Labs, United Kingdom, [e,]University of Exeter, United Kingdom*

**Abstract**

The Jisc-funded UK Research Information Shared Service (UKRISS) project investigated the reporting of research information across the UK HE sector and assessed the feasibility of a national infrastructure based on CERIF with the objective of increasing the efficiency, productivity and reporting quality across the sector. A core reporting profile was developed that would enable harmonised reporting on RCUK-funded research, taking into account the HE-BCI survey as well as REF reporting elements. In this paper we describe the UKRISS modelling approach and provide some insight into the UKRISS reporting objects to support understanding of their formal CERIF representations, i.e. the selection of underlying CERIF entities; the challenges with managing objects and aggregations in CERIF. Example data extracts demonstrate the work.

*Keywords:* UKRISS; CERIF; Research Reporting; Standardisation; Research Results

## 1. Introduction and Background

This paper presents some of the findings of a feasibility and proof-of-concept study into the reporting of research information at a national level within the UK, based on CERIF. The study[1] was carried out by the Jisc-funded UK Research Information Shared Service (UKRISS) project[2].

The reporting of research information is a complex and expensive activity for research organisations. The primary focus of UKRISS was reporting on research projects to the seven government-funded research councils

* Corresponding author. Tel.: +44(0)2082798026.
  *E-mail address:* brigitte.joerg@gmail.com

within the RCUK umbrella organisation. Each council has responsibility for a specific area of research, such as engineering and science (EPSRC) or medical (MRC), and budgets are managed largely autonomously. In the past several years, five of the research councils (AHRC, BBSRC, EPSRC, ESRC, NERC) have implemented an in-house Current Research Information System (CRIS) called the Research Outcomes System (ROS), whereas the remaining two (MRC and STFC) deployed a commercial CRIS called Research Fish. The research information gathered by the RCUK funders is used for a variety of purposes. These include demonstration to government of the impact and value of the funded activities, scoping of future funding programmes and public dissemination.

The data models used across, and even within the reporting systems are substantially different for a number of reasons. At the highest level, the research councils often have their own business objectives for the information fields collected and the way it should be prioritised. There are substantial differences related to the specific disciplines such as the relevance of commercial exploitation of the outputs and differing terminology. Even where there is close alignment of the reporting fields, there are detailed semantic differences that make it difficult to combine and reuse information gathered. Another significant difference between the ROS and Research Fish approaches is that Research Fish relies primarily on direct entry by project PIs, whereas ROS has increasingly moved to harvest information from institutional systems. The data quality of the research information is in some areas a major concern, particularly with the heavy reliance on manual data entry.

As well as the seven government-funded Research Councils (RCUK), a large amount of public research in the UK is funded by charities as well as private organisations. The fields collected relating to research outputs tend to overlap with the information collected by research councils. However, there are also differences. For example, charities tend to place a greater focus on impact narratives that can be used to support future fund-raising.

Research information is also collected approximately every six years in the UK by HEFCE for the large scale national research assessments known as the Research Excellence Framework (REF). HESA also collect for the annual HE-BCI knowledge transfer survey. Both of these were referenced in the harmonisation analysis.

The European Commission is a major funder of research in the UK, but was not in the scope of the current study.

Although UKRISS was focused on research information reporting within the UK, the lack of harmonisation between reporting requirements is also an issue in other countries with multiple funding agencies. For example, Germany has recently launched a project[3] that aims to investigate harmonisation of national reporting. Smaller countries such as Norway[4] and the Czech Republic[5] have less complex funding models and have demonstrated the value and efficiency savings of being able to report and analyse research information at a national level.

The initial remit of the UKRISS project was to assess the feasibility of a national reporting infrastructure. A wide-ranging feasibility study was conducted involving interviews with a representative sample of over forty stakeholders from across the sector including funders, institutions, government bodies, umbrella bodies, CRIS vendors and charity funders. A detailed summary of the finding of the study is described in[1]. The study explored the drivers for harmonisation and requirements for a national reporting infrastructure. The study made three main recommendations:

1. Specification, standardisation and adoption of a core CERIF profile for reporting of research information in UK HEIs.
2. Implementation of a national reporting infrastructure and associated shared services to facilitate the exchange of research information between IT systems within institutions, funders and statutory bodies.
3. Provision of benchmarking tools that enable comparison and analysis of research information generated by multiple organisations for management information purposes.

It is important to make the distinction between reporting services and a reporting system. Stakeholders were not in favour of a national reporting system for a number of reasons. In particular, substantial investments in infrastructure by funders, institutions and government have already been made and there was a wish that any solution should interoperate with these existing systems rather than replacing them.

Recommendation 1 was considered to be a fundamental prerequisite for implementing recommendations 2 and 3. At the same time, there was a need both to demonstrate the feasibility of a national infrastructure as well as some of the potential benefits such as demonstration of enhanced benchmarking and business intelligence functionalities.

With respect to recommendation 1, UKRISS defined the *full profile* as an aggregation of all reporting fields collected by an agreed set of sector bodies such as the research councils. The *core profile* is defined as a set of fields that are common or sufficiently similar in order to be mapped to a single reporting field. The main aim of the

modelling work was to investigate the degree to which harmonisation between initially the RCUK reporting fields could be achieved, to highlight areas where there were significant discrepancies and the reasons for those differences, and where possible to make suitable harmonised definitions.

The Common European Research Information Format (CERIF) was used as the basis for the harmonisation work[6,7]. CERIF has emerged as the preferred format for expressing research information across Europe[8,9,10], and was recommended for adoption by the UK HE sector by the Jisc-funded EXRI-UK report of 2009[11]. Since then CERIF has been piloted for specific applications, but not as a format for reporting requirements across all UK research organisations. Although CERIF allows for a formal description of the UKRISS reporting model, it does not initially provide all the semantics; i.e. object boundary specifications and applicable vocabularies. The UKRISS project has been working closely with the UK chapter of CASRAI, to move towards a standardised set of vocabularies, which are also aligned as far as possible with international conventions[12].

The remainder of the paper is organised as follows. Section 2 contains an analysis of the approach to harmonisation of the RCUK reporting fields and the UKRISS core profile. Section 3 describes validation tools that were developed to demonstrate the value of the core profile for improving data quality. In section 4, we describe the UKRISS Crosswalk connector that provides a lightweight tool for extracting information from multiple institutional IT systems, packaging the information in CERIF format and mapping to a reporting template, which can then be transferred to a funder reporting system. Section 5 presents the lessons learned and section 6 the summary and conclusions.

## 2. The UKRISS Approach

Figure 1 reflects the UKRISS modelling and harmonisation approach starting from the bottom-up with the ROS and ResearchFish models, and continuing top-down with the introduction of an upper reporting level with concepts such as Research Output, Research Transfer, Research Outcome, Research Impact, and Measurement. The upper-level concepts as well as an anticipated use case "institution submits report to funder" guided the development of harmonised reporting objects in the UKRISS profile.
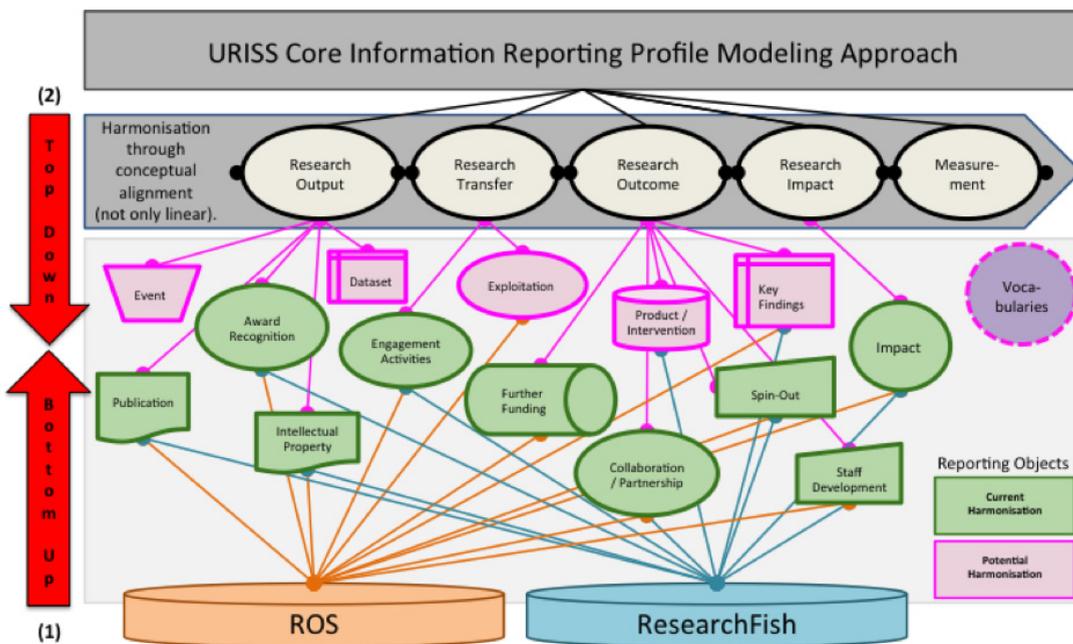


Fig. 1. UKRISS Modelling Approach: bottom-up first / top-down second.

Reporting records are instantiations of UKRISS reporting objects and assignments to upper level reporting concepts do not necessarily follow a linear process or sequence. In practice, a clear or single assignment also varies according to the use case. For example, an institution could view *Staff Development* as a *Measurement* whereas funders view it as a *Research Outcome*. Different and multiple assignments may thus be required or even suggested in support of and depending on multiple stakeholder viewpoints. Figure 1 anticipates a funder's viewpoint.

## 2.1. Harmonisation Analysis

UKRISS defines harmonisation as 'semantic similarity' between comparable entities where entities can be objects, fields or applicable vocabularies. Similarity at field level would for example be given with *Volume* or *Volume Number*, or likewise with *Publication Type* and *Type of Publication* within the boundaries of the reporting object *Publication*. Figure 2 illustrates harmonisation degrees between ROS (blue) and Research Fish (red) reporting objects before (green) and after (purple) the introduction of the UKRISS model. The harmonised UKRISS reporting objects count less string fields and apply a number of relationships instead. Following a CERIF structure this implies the use – and in an optimum case, the re-use – of controlled vocabularies (turquois), and unique identifiers such as e.g. DOI and ORCID.
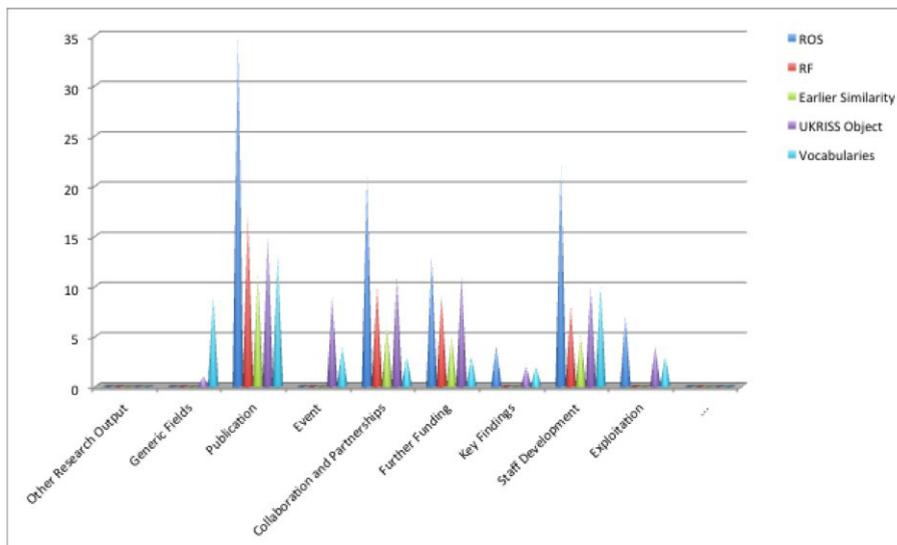


Fig. 2. ROS/RF reporting object harmonisation degrees through UKRISS (for selected objects).

UKRISS investigated inherent and potential degrees of harmonisation within the mentioned UK reporting systems and suggested a gradual approach towards harmonisation guided by conceptual clarity and structural changes in support of sustainability, technological standardisation and use of shared identifiers. In addition to existing reporting objects, UKRISS introduces *Dataset* and *Event* as new reporting objects, and recommends that a future *Impact* object should allow for the recording and time tracking beyond project boundaries. Furthermore *Equipment* and *Education* should be added as future reporting objects to bridge gaps with Higher Education reporting and Research Infrastructures.

Before any implementation UKRISS recommends a clear harmonisation of underlying business requirements across the sector – involving a maximum number of key stakeholders to guide the priorities and the approach towards implementation of proposed extensions – and the rules that apply.

## 2.2. Core Information Reporting Profile

The introduced UKRISS Core Information Reporting Profile upper level reporting concepts were understood as follows (inspired by the ATN approach)[13]:

- *Research Output:* Tangible results describing what was done during research.
- *Research Transfer:* Engagement with end-users during research activity period.
- *Research Outcome:* Changes arising from output; invention or change in approaches to how people behave.
- *Research Impact:* Value-added achieved improvements.
- *Measurement:* Yet to be defined following agreed indicators. Considered important for future comparison amongst data users and data producers.

The entire UKRISS Reporting Profile and its underlying CERIF aggregations are presented in figure 3, where reporting types are subsumed by record types; that is, *Publication* is subsumed under *Research Output*, *Spin-Out* and *Further Funding* is subsumed under *Research Outcome*.
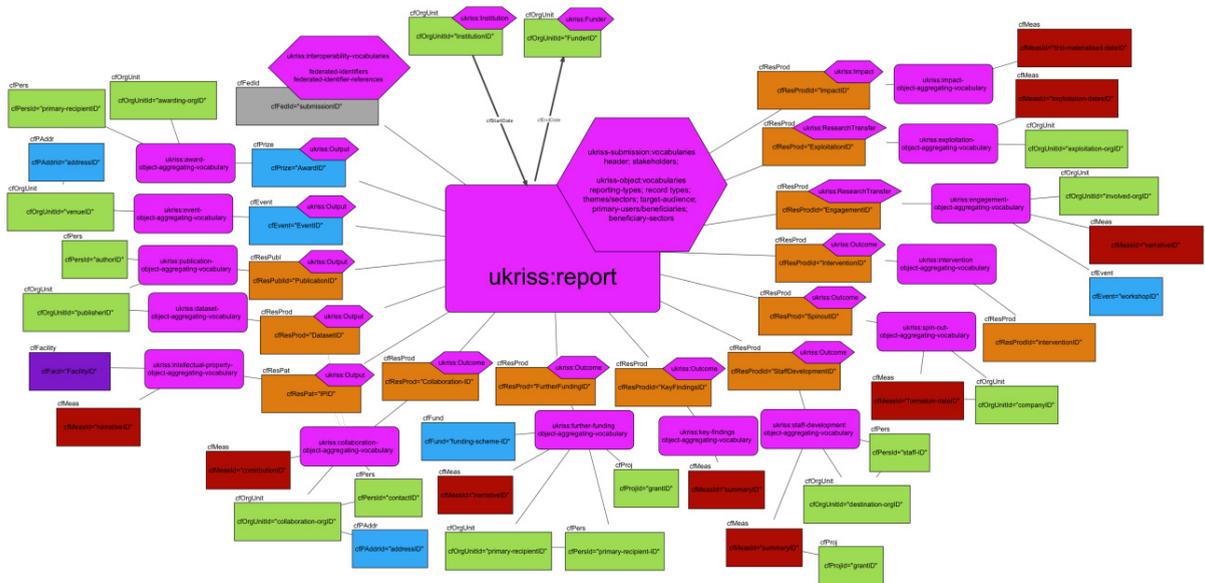


Fig. 3. UKRISS Core Research Information Reporting Profile in CERIF including aggregations
representing the use case "institution submits report to funder".

Figure 3 reflects the use case through the top two CERIF organisation entities *cfOrganisationUnit* with arrows showing the information flow through a report object – an aggregation of multiple CERIF entities and vocabularies – from an institution to a funder. This ukriss:report object in figure 3 aggregates all UKRISS reporting objects and vocabularies, while each UKRISS reporting object is itself an aggregation of CERIF entities and vocabularies.

The entire UKRISS Core Information Reporting Profile is available in an Excel sheet from the UKRISS project website: http://ukriss.cerch.kcl.ac.uk/, with CERIF mappings for each reporting object. More detailed documentation is provided with the final UKRISS report and CERIF XML example files for each UKRISS reporting object[14]. Where *Research Output* objects such as *Publication*, *Dataset*, *Intellectual Property*, *Event*, or *Award* are explicit entities in the CERIF model – namely *cfResultPublication*, *cfResultProduct*, *cfResultPatent*, *cfEvent*, *cfPrize* – this is not the case with *Research Transfer* objects or *Research Outcome* objects, which are modelled as *cfResultProduct* in a first instance, while aggregating functionally related entities in a second instance. For example, *Staff Development* is at first considered an outcome *cfResultProduct* and secondly a person and its aggregates. The same holds for *Collaboration* at first considered an outcome, *cfResultProduct* and secondly an organisation and its aggregates (see also figures 1 and 3).

*2.3. Example UKRISS Reporting Object Extracts*

The UKRISS model anticipates the recommended harmonisation changes. Some example extracts of reporting objects in CERIF XML illustrate the underlying model and its implementation.



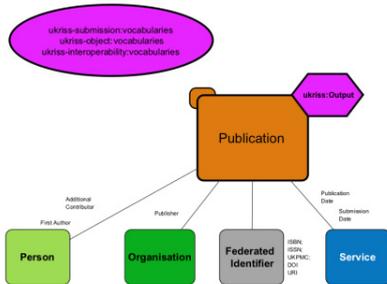Fig. 4. Aggregation of UKRISS Publication elements.

```
<cfResPubl>
 <cfResPublId>3-FA88911-722B-4B05-B5FF-590926D</cfResPublId>
 <cfResPublDate>2012-10-01</cfResPublDate>
 <cfTitle cfLangCode="en" cfTrans="o">Sonic prosthetics:
   Exploring posthuman personal space in digit ..</cfTitle>
   …
 <cfFedId>10.1386/padm.7.2.155_1</cfFedId>
 <cfFedId_Class>
  <cfClassId>cerif:doi-uuid</cfClassId>
  <cfClassSchemeId>ukriss-object:federated-identifiers-uuid</cfClassSchemeId>
 </cfFedId_Class>
 <cfResPubl_Class>
  <cfClassId>ukriss-object:publication-uuid</cfClassId>
  <cfClassSchemeId>ukriss:record-types-uuid</cfClassSchemeId>
   …
</cfResPubl>
```



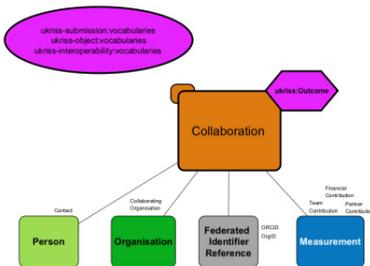Fig. 5. Aggregation of UKRISS Collaboration elements.

```
<cfResProd>
 <cfResProdId>3--104514</cfResProdId>
 <cfName cfLangCode="en" cfTrans="o">Role of the T-type calcium channel
      in cardiac myocyte hypertrophy ...</cfName>
   …
 <cfResProd_Class>
  <cfClassId>ukriss-object:collaboation-uuid</cfClassId>
  <cfClassSchemeId>ukriss:record-types-uuid</cfClassSchemeId>
 </cfResProd_Class>
 <cfResProd_Class>
  <cfClassId>ukriss:outcome-uuid</cfClassId>
  <cfClassSchemeId>ukriss:reporting-types-uuid</cfClassSchemeId>
</cfResProd_Class>
   …
</cfResProd>
```

The modelling of the UKRISS reporting objects did not require any extensions to the CERIF model. However, it encouraged consideration of a more standardised / formal approach with respect to CERIF object boundary definitions and their application according to specified requirements or 'business rules'. These will additionally support any related tool implementation and data validation.

## 3. UKRISS Validation, Visualisation and Aggregation Tools

UKRISS developed a number of practical proof-of-concept tools based on the core profile. These tools demonstrate the wider range and power of harmonised models to support research information interchange and reuse.

Harmonisation has the potential to improve data quality, enhance understanding through visualisation tools and enable advanced analytics through data aggregation.

Validation is the process of ensuring compliance with the model and the quality of the data represented therein. Having a well-understood document format, with known data types in known fields (e.g. knowing when a field is supposed to be a date, and even what format the date should be represented in) means that generic validation software can be developed, which can then be used by and to the benefit of an organisation that works with the information. During the course of the UKRISS project a tool was developed to validate individual elements of the model (such as checking that an ISSN is an ISSN). It can also look up data in external data-sources and cross-reference with the document being validated in support of consistency with the metadata across the community.
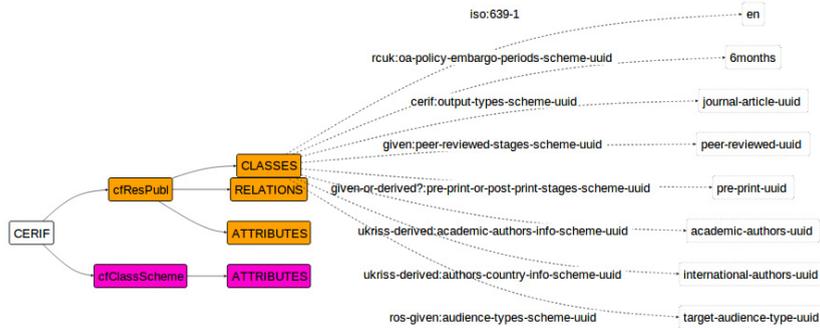
Validated data records can then be visualised according to the UKRISS model based on formal CERIF (see figure 6). A well-structured and a well-understood data model is also a pre-requisite for quality data aggregation.



Fig. 6. Visualisation of a validated Publication record in CERIF.

## 4. UKRISS Crosswalk Connector

An open source connecter has been developed to extract research data from existing institutional systems, transform the data into the UKRISS CERIF format, and load the data securely to a location accessible by the external data recipient, e.g. the funding body as shown in figure 7.

The guiding development principles were that it should be free to download, with a low barrier to entry in that it should be easy to install and intuitive to use without need for extensive technical resources or expertise, and thus be valuable not only to research intensive institutions, but also to smaller organisations.
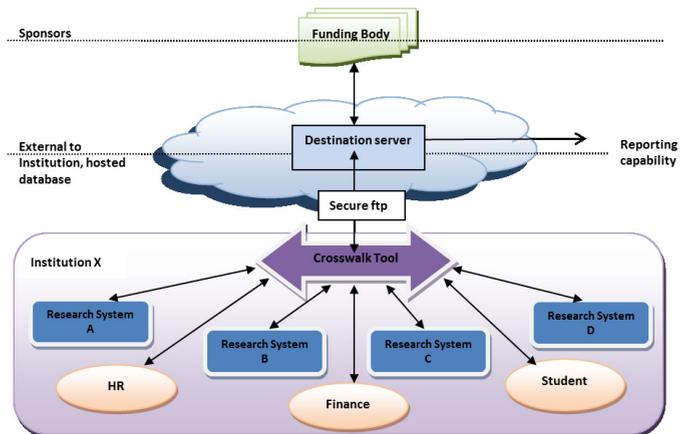


Fig. 7. Schematic of the Crosswalk Connector interfaces.

One of the principal benefits of aggregating data from multiple sources derives from how effective this process is at highlighting errors and inconsistencies with source data. More information is available in the final report[14].

## 5. Lessons Learned

We have demonstrated the feasibility and potential benefits of harmonised research reporting guided by a formal domain model for the multiple involved organisations in the research ecosystem. Such a model ensures the

consistent usage and re-usage of defined reporting objects and sub-objects. All stakeholders need to participate in the definition of requirements to ensure a conceptual clarity and as well support with the process of implementation, if harmonisation is to benefit the national, and subsequently the global research ecosystem.

## 6. Summary and Conclusion

Harmonised national research reporting can only be achieved through agreement between the stakeholders involved. Clearly defined requirements are the pre-requisite for sustainable conceptual descriptions and enable a consistent and valid implementation nationally and internationally. CERIF guided the harmonisation process by providing a research domain model that allows for the formal representation of any context, profile or object. However, CERIF needs to dedicate further thoughts (such as rules) towards formal definitions of contextual or object boundaries and their applicable vocabularies.

## Acknowledgements

## References

1. Waddington S, Sudlow A, Walshe K, Scoble R, Mitchel L, Jones R, Trowell S. Feasibility Study into the Reporting of Research Information at a National Level Within the UK Higher Education Sector. *New Review of Information Networking 2013,* Vol 18, Issue 2, pp. 74-105.
2. Waddington S. UKRISS Proposal within Jisc RIM Programme: http://ukriss.cerch.kcl.ac.uk/wp-content/uploads/2012/05/JISC-RIM-proposal-Final_PUBLIC.pdf  (see also UKRISS blog: http://ukriss.cerch.kcl.ac.uk)
3. German Science Council (Wissenschaftsrat): Empfehlungen zu einem Kerndatensatz Forschung: http://www.wissenschaftsrat.de/download/archiv/2855-13.pdf (see also project website Research Core Dataset: http://www.research-information.de/Projekte/Research_Core_Dataset/projekte_research_core_dataset.asp)
4. Sidselrud A, Lingjærde GC. The practical implementation of the CRIS system CRIStin and the goals/challenges of bringing 150 institutions into production within a year. In: Jeffery K, Dvořák (eds.): *E-Infrastructures for Resarch and Innovation: Linking Information Systems to Improve Scientific Knowledge Production.* Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic), pp. 305-312. ISBN 978-80-86742-33-5.
5. Chudlarský T, Dvořák J. A National CRIS Infrastructure as the Cornerstone of Transparency in the Research Domain. In: Jeffery, Keith G; Dvořák, Jan (eds.): *E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production*: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic), pp. 9-17. ISBN 978-80-86742-33-5.
6. CERIF XML 1.6 Release: http://www.eurocris.org/Index.php?page=CERIF-1.6&t=1
7. Jörg B. CERIF: The Common European Reseach Information Format Model. Data Science Journal. Volume 9, Special Issue: CRISs for the European e-Infrastructure (Jul. 2010), CRIS24-31.
8. Houssos N, Jörg B, Matthews B. *A Multi-Level Metadata Approach for a Public Sector Information Data Infrastructure.* In: Jeffery, Keith G; Dvořák, Jan (eds.): *E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production*: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic)
9. OpenAIRE Guidelines for CRIS Managers: https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_CRIS
10. Houssos N, Jörg B, Dvorak J, Príncipe P, Rodrigues E, Manghi P, Karstensen Elbæk M. *OpenAIRE Guidelines for CRIS Managers: Supporting Interoperability of Open Research Information through established standards*. Procedia Computer Science. CRIS 2014. May 13-15, 2014
11. Rogers N, Huxley L, Ferguson N. *Exchanging Research Information in the UK. EXRI-UK: A study funded by JISC.* 2009. Web. 6. Sept. 2013. http://repository.jisc.ac.uk/448
12. Baker D. Solving data disconnect in global research administration: the CASRAI approach. Research Gloabal, November 2013
13. Duryea M, Hochman M, Parfitt A: Measuring the impact of Research. Research Global, February 2007. http://www.atn.edu.au/Documents/Articles/2011/2010/2009/2008/2007/Measuring%20the%20impact%20of%20research.pdf.
14. Final UKRISS project report available from the UKRISS blog: http://ukriss.cerch.kcl.ac.uk