UK Research Information Shared Service Project

(UKRISS)

Final Report

Jisc Final Report

Project Information			
Project Identifier	*To be completed by Jisc*		
Project Title	UK Research Information Shared Service (UKRISS)		
Project Hashtag	jisc_ukriss		
Start Date	1 March 2012	**End Date**	31 January 2014
Lead Institution	King's College London		
Project Manager	Simon Waddington		
Contact email	simon.waddington@kcl.ac.uk		
Partner Institutions	British Library, Brunel University, Cottage Labs, University of Exeter, euroCRIS, University of Edinburgh (unfunded) *Subcontractors:* Viewforth Consulting, Certus Technology Associates		
Project Web URL	http://ukriss.cerch.kcl.ac.uk		
Programme Name	*Research Information Management*		
Programme Manager	Neil Jacobs, Verena Weigert		

Document Information			
Author(s)	Brigitte Joerg, Richard Jones, Diane McDonald, Richard Gartner, Monique Ritchie, Rosa Scoble, Allan Sudlow, Emanuil Tolev, Stephen Trowell, Simon Waddington.		
Date	4 July 2014	**Filename**	
URL	http://ukriss.cerch.kcl.ac.uk		
Access	This report is for general dissemination		

Executive summary

The reporting of research information is a complex and expensive activity for research organisations (ROs). There is little alignment between funders of the reporting requests made to institutions and requests made to individual researchers about their research outputs and outcomes. This inevitably results in duplication and increased costs across the sector, whilst limiting the potential sharing and reuse of the information.

The UK Research Information Shared Service (UKRISS) project conducted a feasibility and scoping study for the reporting of research information at a national level based on CERIF (Common European Research Information Format), with the objective of increasing efficiency, productivity and quality across the sector. The aim was to define and prototype solutions which are compelling, easy to use, have a low entry barrier, and support innovative information sharing and benchmarking.

CERIF has emerged as the preferred format for expressing research information across Europe. To date, CERIF has been piloted for specific applications, but not as a format for reporting requirements across all UK ROs.

The aim of this report is to present the work carried out by the UKRISS project, including requirements gathering, modelling and prototyping, as well as recommendation for sustainability. UKRISS was divided into two phases. Phase 1, mapping the reporting landscape, ran from March 2012 to December 2012. Phase 2, exploring delivery of potential solutions, began in February 2013 and ended in December 2013.

Feasibility study

In phase 1, an extensive study to determine the options for research reporting at a national level was carried out. This involved a wide range of stakeholders including research organisations, research councils, statutory bodies and charities that fund research. The study comprised a formal feasibility study, based on a study of research information reporting use cases and business needs.

In general, stakeholders reported many common issues. A set of drivers and requirements for harmonisation were extracted. Drivers for harmonisation were coalesced around six themes:

D1. Improve business intelligence, management and due diligence through better information quality and reporting utility.
D2. Reduce the reporting burden and increase the efficiency or response agility of the research community through harmonisation of reporting processes and/or systems.
D3. Enable cross-sector impact analysis, evaluation and strategy development through systemisation and harmonisation of reporting.
D4. Increase research community reporting compliance through deploying easy-to-use flexible reporting systems with user benefits.
D5. Improve the research, strategy and planning across UK institutions through use of better quality reporting information.
D6. Improve research information management across the sector through deploying sustainable, affordable solutions that are fit-for-purpose.

In parallel, specific requirements for harmonisation were extracted as follows:

R1. Harmonise dictionaries and usage of CERIF within the UK HE sector.

R2. Obtain agreement between all key stakeholders (eg funders, institutions, charities, statutory bodies) on closer alignment of reporting requirements and their persistence, and adoption.
R3. Provide structures (common Application Programming Interfaces (APIs), shared services or connectors) to support the exchange of research information, but not a central reporting system.
R4. Increase the quality and timeliness of research information across the sector.
R5. Facilitate the flow of information between internal institutional systems and external systems (eg funder systems) in CERIF format.
R6. Enable institutions to more effectively consume and reuse research information (eg for benchmarking and management information, portfolio management, collaboration, compliance monitoring, communications).
R7. Support benchmarking and portfolio analysis across research funders.
R8. Provide appropriate data governance, transparency and security when collecting, sharing and reusing sensitive research information.

The study uncovered a fragmented research information reporting landscape, with a lack of alignment in information requests made to institutions and researchers. Information is often collected in similar but different ways, placing a burden on researchers and research administrators in institutions. This diversity was also reflected in a lack of automation of key reporting processes. Low data quality was a major issue in some areas. Where this occurred, in part it was due to low reporting compliance, resulting in incomplete data. In other cases, errors were caused by manual processing and lack of cross-referencing of data stored across different systems. All of these issues limited the sharing and reuse of research information for purposes such as business intelligence, strategic planning and benchmarking.

Based on this analysis, the feasibility study identified three specific areas for further work:

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. 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. Much of the cost of integration with a national system would fall on institutions, and these costs were perceived as being potentially high. There was however support for common services amongst a wide range of stakeholders. In particular, funders were keen to have common API definitions for systems to support automated harvesting of reporting information.

Technical results

Interoperability at a semantic level was the main focus for phase 2, as this was seen as fundamental to developing national-level reporting services and business intelligence tools. However, given the abstract nature of modelling and the need to convince key stakeholders of the benefits and value of a core profile, supporting work on technical demonstrators for data exchange, visualisation, benchmarking and aggregation were also produced and evaluated.

Core profile

A detailed bottom-up analysis of reporting objects in the RCUK outcomes systems ROS and Research Fish uncovered a "semantic gap" between their object describing fields. In order to resolve this problem, a top-down approach was taken, resulting in the introduction of an upper level of reporting concepts labelled as Research Output, Research Transfer, Research Outcome, Research Impact, and Measurement, under which reporting objects were subsumed. These concepts were used as a basis for aligning the reporting objects and fields at a lower level.

UKRISS defined harmonisation as "semantic similarity" between compared entities, where entities can be objects, fields, or applicable vocabularies. We then estimated the actual and potential degree of harmonisation that could be achieved, taking into account also the HEFCE HE-BCI survey and RE reporting objects. There is an opportunity to produce about a 30% harmonisation between ROS and Research Fish by aligning the definition of similar fields and merging of vocabularies. There is a 50% discrepancy between the information requests made by councils to institutions that cannot be resolved by minor adjustments of existing fields. This is due in part to the different business needs of the councils and discipline-specific reporting requirements.

Validation, visualisation and aggregation tools

Proof-of-concept tools for applying the UKRISS core profile were developed in three areas:

- Validation is the process of ensuring compliance with the model and the quality of the data represented. The tool developed in UKRISS can validate individual elements of the model (such as checking that an ISSN is an ISSN), but it can also look-up data in external data-sources (such as Entrez) and cross-reference that with the document being validated
- Visualisation is the process of producing a visual representation of the core model that enables users to explore their data more easily. The software produces a collapsible-tree representation of CERIF-XML and, although quite basic, shows the potential for developing more comprehensive CERIF dashboards
- Aggregation is the process of bringing together and exploring many objects which are all expressed using the model. In this section, we implemented a specific use case gathered by the Jisc-funded G4HE project to demonstrate how the model facilitates analysis and mining of information by virtue of the harmonised representation

The work highlighted the considerable opportunities for further development, particularly in the areas of validation and aggregation tools.

Crosswalk Connector

The UKRISS Crosswalk Connector was developed to show how data can be readily extracted from existing institutional systems to populate funders' reports, eg ROS/Research Fish. The Connector is open source software that extracts research data from institutional systems, transforms the data into the UKRISS CERIF profile, and loads the data securely to a location accessible by the external data recipient, eg funding body. The guiding principles for the Connector 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. This enables the Connector to be valuable not only to research-intensive institutions, but also smaller organisations that may not have the same degree of technical capability. The solution accommodates a range of data types and is therefore compatible with a wide range of existing systems (and spreadsheets). One of the principal benefits of aggregating data from multiple sources derives from how effective this process is at highlighting errors and inconsistencies in the source data. Significant improvements in source data quality arise from implementation of this type of approach.

Preparatory business case

The scope of the preparatory business case work was to describe, qualitatively, for each stakeholder group, the benefits of harmonisation (including negative benefits), the risks and their potential impact, the types of costs involved and the barriers to uptake. The work aimed to lay the groundwork for a more comprehensive quantitative study, once RCUK plans for research information are known.

We concluded that for all the stakeholders the benefits of harmonisation are likely to outweigh the dis-benefits. However, there are significant dis-benefits to be addressed, particular around costs that might put small institutions and charities at a disadvantage. Although there are risks associated with implementing harmonisation, these were outweighed by the risk of not doing so. There was also a significant risk that a partially harmonised model developed by the research councils is adopted, which may not meet the business and reporting needs of the wider HE community.

The business case for harmonisation of research reporting is situated within a complex innovation ecosystem involving multiple stakeholders. The study considered each of the key stakeholders as well as the potential contribution to increasing economic and social benefits to business and society.

Recommendations

As a result of the work performed, the project has made a number of recommendations for stakeholders involved in research information reporting, the wider HE community and for Jisc.

Specific recommendations

1. Promote the standardisation of the core reporting profile and vocabularies through the CASRAI UK group or another forum with cross-sector representation *[RCUK, institutions, charities and statutory bodies]*.
2. Investigate the closer business alignment of reporting fields in terms of their value versus the effort required to collect them *[RCUK supported by institutions and statutory bodies].*
3. Promote alignment between teaching and research information, particularly in the development of sector-wide vocabularies, and the area of research students, interacting with HEDIIP *[RCUK, statutory bodies].*

Recommendations for the wider community

1. Consider the establishment a single organisation for governance of the core profile.
2. Consider adoption of the core profile (or portions of it) by non-RCUK funders and statutory bodies. In particular, the UKRISS core profile should be considered as an important element of future Research Excellence Frameworks (REFs) *[other funders including charities, statutory bodies].*

Recommendations for Jisc

1. Arrange a meeting of senior representatives of the key stakeholders, including RCUK funders and institutions to review and follow up the UKRISS findings and to carry forward the proposals for harmonisation.
2. Promote the development of the core profile through support for the CASRAI UK task groups and other initiatives. This includes continuing support for CASRAI over a timeframe that enables consensus to be reached.
3. Promote the development of an open ecosystem of research information systems based on CERIF by facilitating

engagement between stakeholders and funding of further pilots and studies.

4. Promote the prototyping and adoption of registries for identifiers to improve data quality and interoperability.
5. Conduct a more comprehensive cost-benefit analysis to build on the benefits, risk and measurements identified in the UKRISS preparation for a full business case for harmonisation. A prerequisite would be buy-in from both funders and institutions.
6. Investigate the applicability of UKRISS techniques to address analogous issues in the return of teaching and other non-research information from institutions to external bodies.

Contents

1. Acknowledgements

The project team would like to thank Jisc for funding UKRISS as part of the Research Information Management programme.

The project received a considerable input from the UKRISS Steering Board, in terms of planning the work, giving expert input, providing useful contacts and introductions, as well as reviewing and improving the deliverables. We would like to thank all the Steering Board members Ian Carter, Maja Maricevic, Liz Philpots, Geraldine Clement-Stoneham, Gerry Lawson, Luke Moody, Geoff Rodgers, Neil Jacobs, Kimberley Hackett, Kevin Dolby, Ian McArdle, Luke Taylor, Verena Weigert and Gregor McDonagh for their contributions and in particular Ian Carter for his role as chair. The full list of Steering Board members with their affiliations is included in Appendix 1. In addition to the Steering Board members, we would particularly like to thank Beverley Sherbon (MRC), Gavin Reddick (MRC) and Ben Ryan (EPSRC, CIS) for their input and guidance during the project.

In phase 1 of UKRISS, the project team interviewed over 40 key stakeholders in research information reporting across the sector. We would like to thank all those people who consented to being interviewed for their time and inputs, which contributed greatly to providing a comprehensive and balanced view of the current landscape.

The UKRISS project held two final workshops in London and in Glasgow. We would like to thank the British Library for hosting the London event, and Valerie McCutcheon for inviting the project team to run a second workshop at the University of Glasgow Library.

2. Project summary

The reporting of research information is a complex and expensive activity for research organisations (ROs). There is little alignment between funders of the reporting requests made to institutions and individual researchers about their research outputs and outcomes. This inevitably results in duplication and increased costs across the sector, whilst limiting the potential sharing and reuse of the information.

The UKRISS project conducted a feasibility and scoping study for the reporting of research information at a national level based on CERIF, with the objective of increasing efficiency, productivity and quality across the sector. The aim was to define and prototype solutions which are compelling, easy to use, have a low entry barrier, and support innovative information sharing and benchmarking.

The Common European Research Information Format (CERIF) (2014) has emerged as the preferred format for expressing research information across Europe. CERIF has been piloted for specific applications, but not as a format for reporting requirements across all UK ROs.

The UKRISS project was divided into two phases. Phase 1 ran from March 2012 to December 2012. Phase 2 began in February 2013 and ended in December 2013.

In phase 1, an extensive study to determine the options for research reporting at a national level was carried out. This involved a wide range of stakeholders including research organisations, research councils, statutory bodies and charities that fund research. The study took the form of a formal feasibility study, based on a study of research information reporting use cases and business needs across the sector, analysing their strengths and drawbacks, dependencies and risks and setting out recommendations for further work.

The study made three main recommendations:

1. Given the scale of these tasks, the main focus of phase 2 was on recommendation 1, which was regarded as a key enabler for future developments. 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.

In order to provide compelling evidence to support recommendation 1, the project also worked on prototypes and demonstrators related to 2 and 3, as well as preliminary work on the business case for harmonisation of research reporting outcomes.

Specific outputs from phase 2 of UKRISS were:

- A detailed analysis of the reporting fields in the RCUK Research Outcomes System (ROS) and Research Fish together with the HESA HE-BCI survey, in order to determine the opportunities for harmonisation
- A conceptual framework for harmonisation, and definition of aggregated and core reporting profiles for RCUK and HE-BCI reporting including mappings of the fields to CERIF and collection of common vocabularies
- An analysis of the benefits, risks and issues relating to harmonisation of research information reporting and the adoption of the UKRISS core profile in preparation for the development of a full business case
- A set of software prototypes including data validation for the core profile, business analytics tools to exploit aggregated research information and CERIF visualisation tools
- The UKRISS Crosswalk Connector, which partially automates the process of extracting reporting information from institutional systems and mapping it to the appropriate reporting profile for submission to funders

The UKRISS core profile, CERIF mappings and associated vocabularies will be submitted to the CASRAI Contributions and Open Access Working Group for consideration as a standard, CASRAI (2014). Further information on the project and the full set of outputs are available on the project blog at http://ukriss.cerch.kcl.ac.uk.

3. Main body of report

This section describes the main outputs, outcomes, impact and lessons learned from the UKRISS project. The main focus of the report is on the work completed in phase 2. However, for the benefit of the reader, we include a summary of the main phase 1 findings.

3.1 Project outputs and outcomes

In this section, the outputs and outcomes of the UKRISS project are described. Where possible, URLs are included to the relevant resources. For the convenience of the reader, the major project deliverables are listed in section 3.1.1 and a more detailed list of outputs and outcomes is contained in section 3.1.2.

3.1.1 Major deliverables

Table 3.1 describes the main project deliverables. Unless stated otherwise, documents and other project outputs will be made available through the project blog at http://ukriss.cerch.kcl.ac.uk/deliverables.

Output / Outcome Type	Phase	Brief Description and URLs (where applicable)
Report	1	Project plan
Report	1	Feasibility study report, describing the requirements for harmonisation of reporting and the recommendations for further development
Journal paper	1	Feasibility Study Into the Reporting of Research Information at a National Level Within the UK Higher Education Sector. *New Review of Information Networking* Volume 18, Issue 2, 2013, pp.74-105. DOI:10.1080/13614576.2013.841446 www.tandfonline.com/eprint/76QMhvSWJTEZMcghVewh/full#.Unt_wV9FCUk
Report	2	Phase 2 project plan
Software	2	UKRISS model validation tool http://cottagelabs.com/news/metatool-making-metadata-better-data
Software	2	CERIF schema visualisation tool
Software	2	Aggregation tools http://cottagelabs.com/news/an-academic-catalogue
Software	2	Crosswalk Connector Documentation: http://certus-tech.github.io/crosswalk-connector-doc/en-US/Crosswalk/0.1/html/UKRISSCrosswalkTool/index.html Code and source files: Java source, documentation source and build scripts https://github.com/certus-tech/crosswalk-connector-src Kettle plugins, README for installation instructions: https://github.com/certus-tech/crosswalk-connector-bin Documentation build: https://github.com/certus-tech/crosswalk-connector-doc

Report	2	Final report
Model	2	Core Information Reporting Profile in CERIF Final report – Appendix 4 (CERIF XML examples).
Website	1 & 2	UKRISS blog http://ukriss.cerch.kcl.ac.uk

Table 3.1 Main project deliverables

3.1.2 Detailed list of outputs and outcomes

Table 3.2 contains a more comprehensive list of all relevant project outputs and outcomes.

Output / Outcome Type	Phase	Brief Description and URLs (where applicable)
Report	1	Dissemination plan (project and Jisc internal)
Report	1	Landscape study http://ukriss.cerch.kcl.ac.uk/?p=75
Report appendix	1	Feasibility Study Appendix B: Stakeholder analysis describing the main players in research information reporting and their interests and influence on harmonisation. Appendices
Report appendix	1	Feasibility Study Appendix D: Technology review, describing the most relevant available technologies in the research information management area
Report appendix	1	Feasibility Study Appendix G: Drivers for harmonisation
Report appendix	1	Feasibility Study Appendix H: Full requirements list for harmonisation
Report appendix	1	Feasibility Study Appendix K: Summarised requirements for harmonisation
Conference paper	1	CRIS2012 conference poster paper http://ukriss.cerch.kcl.ac.uk/ukriss-poster-at-cris-2012
Presentation	1	UCISA-CISG conference presentation: UKRISS: reporting and exchange of research information at a national level

		www.ucisa.ac.uk/groups/cisg/Events/2012/cisg2012/Programme.aspx
Poster and blog post	1	ARMA conference poster http://ukriss.cerch.kcl.ac.uk/arma-conference-poster
Report chapter	1	Impact Information Management Systems. In Dean *et al* (Eds) (2013) 7 Essays on Impact. DESCRIBE Project Report for Jisc. University of Exeter. www.exeter.ac.uk/media/universityofexeter/research/inspiringresearch/describeproject/pdfs/2013_06_04_7_Essays_on_Impact_FINAL.pdf
Report	2	Contribution to HEDIIP report The HE Information Landscape: Creating and Managing a Data Model, July 2013 www.hediip.ac.uk/wp-content/uploads/HEDIIP_Data_Language_Report_2013-09.pdf
Report appendix	2	ROS, Research Fish and HE-BCI reporting field analysis, CERIF mappings and vocabularies and CERIF XML data examples Appendix 4 (CERIF XML) to the final report
Spreadsheet and blog post	2	CERIF elements and vocabularies landscape survey http://ukriss.cerch.kcl.ac.uk/cerif-elements-and-vocabularies-landscape-survey
UKRISS workshop	2	UKRISS London workshop (01/11/2013) http://ukriss.cerch.kcl.ac.uk/ukriss-london-workshop
UKRISS workshop	2	UKRISS Glasgow workshop (20/11/2013) http://ukriss.cerch.kcl.ac.uk/ukriss-glasgow-workshop
Presentation and blog post	2	UKRISS presentation at Reconnect13, Canada. Alignment of UKRISS with international reporting harmonisation activities http://ukriss.cerch.kcl.ac.uk/ukriss-at-reconnect13
Report appendix	2	Business case for harmonisation – Appendix 7 to final report
Conference paper	2	Paper submitted to CRIS2014 conference. www.cris2014.org
Presentation	1	UKRISS at German International Forum www.forschungsinfo.de/kerndatensatz/en/index.php?mitteilungen (Jan 2014)

Presentation	2	Presentation on UKRISS findings to CASRAI Reporting and OA group (20/11/2013)

Table 3.2: Detailed list of deliverables

3.2 How did you go about achieving your outputs/outcomes?

This section provides a summary of the aims, objectives and methodology adopted in UKRISS.

3.2.1 Project structure and initial objectives

The initial broad aims and objectives of UKRISS were to:

- Explore the feasibility and proof-of-concept delivery of a national shared service for reporting of research information
- Increase the flow of research information around the sector whilst reducing the burden of research reporting on individual researchers and institutions

The initial broad objectives were refined during the course of the project as described in the following sections.

3.2.2 Phase 1 overview

Landscape study

The landscape around reporting of research information is complex. There are many ongoing initiatives, both at a national level as well as internationally. The project carried out an extensive landscape study covering existing reporting systems, projects and standards. The UKRISS Landscape Study (2013) covered three main aspects:

- Collation of a list of items containing current and recent projects, developments in the sector and systems for the exchange and reporting of research information
- Classification of key characteristics for each item in the landscape list to develop a "landscape study" taxonomy
- Use of the taxonomy for the identification of the most significant activities to help develop the integration framework for UKRISS

The landscape study identified 60 key items that were of direct relevance to the work on UKRISS. These included:

1. Work of euroCRIS on the development of CERIF, the EXRI-UK report, Rogers *et al* (2009) that recommended the adoption of CERIF within the UK HE sector and the supporting business case report, Bolton (2010).
2. Projects in the Jisc Research Information Management (RIM) programme, including RMAS, IRIOS 1 and 2, CERIF in Action, the BRUCE Project (2012), Readiness for REF and DESCRIBE.
3. Jisc projects in the repositories area, including RiO Extension Project (2012).
4. Identifier initiatives including ORCID (2013), FundRef (2013), GrantRef (see CrossRef, 2013) and Ringgold (2014).
5. RCUK grant and outputs management systems, including the Je-S portal, Research Fish and Research Outcomes System (ROS).

6. CRIS systems used by institutions such as Elsevier Pure (2013), Thomson Reuters Converis (2013) and Symplectic Elements (2013), as well as the ePrints repository system (2013).
7. The BIS Gateway to Research (GtR) project (2013).

Stakeholder analysis

Informed by the landscape study, a comprehensive list of relevant individuals, research organisations, funding bodies and professional associations representing key stakeholders in the domain was generated. The rationale for scoping down this long list to a feasible number of stakeholders to approach was based on who had deployed, funded or was using a system to support research information management. The identified stakeholders were categorised into broad typologies based on role type in relation to the project. Stakeholders were then stratified within these broad typologies to ensure a representative sample of organisations across sectors of different sizes and maturity. The composition of the sample group is described in Table 3.3.

Category	Typology	Description	Number of subjects
Funders	Funder	Government-backed funder (eg RCUK)	5
	Charity	Charity funder	4
HE Organisations	HE Organisation	(eg HEFCE, HESA)	2
Institutions	GuildHE	Institutional grouping	2
	Alliance	Institutional grouping	2
	Million plus	Institutional grouping	2
	1994	Institutional grouping	2
	Russell	Institutional grouping	2
	Research Institutes	eg British Library	2
Researchers	Researcher	Researcher at institution	2
Umbrella Organisations	Umbrella	Umbrella (eg ARMA, UCISA)	3
Vendors	Vendor	CRIS vendor	3

Table 3.3: Stakeholder typologies and interviews

Research funders have traditionally determined the reporting requirements for their awards, and thus were seen as key stakeholders. Institutions were also seen as important, although the requirements of large intensive research institutions were perceived to be very different to institutions with only a small number of grants. Therefore institutions were subdivided into multiple categories. The main focus of the institutional approaches was on senior research managers and administrators. We also approached a small number of project Principal Investigators to understand in detail the workload and difficulties of generating reports. National bodies such as HEFCE and HESA were also seen as important, although with reference to the HESA Information Landscape Study (2013), they do not have an interest in imposing solutions on institutions and funders. HEFCE and HESA have an interest in gathering data from the sector in exercises such as REF and the HE-BCI survey, which overlaps with information gathered by funders. Vendors were seen as less critical in shaping harmonisation, but they had an important role and interest in implementing any such solution.

Requirements study

A comprehensive set of 64 interview questions was produced, mapped to typologies, as well as functional and non-functional requirements, to maximise the utility of the qualitative information captured for translation into requirements across different types of stakeholder. These questions were tailored before each interview to fit the roles and responsibilities of the interviewee. The questions covered such areas as the objectives of UKRISS, existing processes and systems, reuse of research information, exchange of information with external systems and ease of use.

A representative sample of over 40 stakeholders was approached, with each interview lasting approximately one hour. Audio recordings were transcribed to text, and raw requirements extracted. The requirements were clustered and de-duplicated according to a two-level taxonomy. In parallel a set of key drivers were extracted. Use cases related to reporting, exchange and reuse of research information were identified.

Drivers for harmonisation

Drivers were filtered according to these categories, analysed for common themes, de-duplicated and coalesced around six overarching driver descriptions sharing a common format:

[overall aim] *through* **[improvement].**

The main drivers extracted were:

D1. Improve business intelligence, management and due diligence through better information quality and reporting utility.
D2. Reduce the reporting burden and increase the efficiency or response agility of the research community through harmonisation of reporting processes and/or systems.
D3. Enable cross-sector impact analysis, evaluation and strategy development through systemisation and harmonisation of reporting.
D4. Increase research community reporting compliance through deploying easy-to-use flexible reporting systems with user benefits.
D5. Improve the research, strategy and planning across UK institutions through use of better quality reporting information.
D6. Improve research information management across the sector through deploying sustainable, affordable solutions that are fit-for-purpose.

Requirements relating to reporting harmonisation

Eight main requirements were extracted from the UKRISS study. Each requirement has a high level description supplemented by a more detailed set of sub-requirements.

R1. Harmonise dictionaries and usage of CERIF within the UK HE sector.

   a. Produce a common set of definitions of data dictionaries, output types including non-publications, identifiers (people, equipment, grants, and funders), institutional structures, research topics and metrics.
   b. Specify use of DOIs for linking outputs and equipment to grants and funders, outputs to researchers etc.
   c. Align more closely standards development and implementation with the practical requirements of a wide range of stakeholders.
   d. Support international initiatives such as ORCID (2013), FundRef (2013) and CrossRef (2013).

R2. Obtain agreement between all key stakeholders (eg funders, institutions, charities, statutory bodies) on closer alignment of reporting requirements and their persistence, and adoption.

   a. Define a minimum core dataset that is collected by all stakeholders to enable comparison, sharing and reuse.
   b. Enable reporting information to be collected once and associated to multiple funders.
   c. Develop agreed definitions of non-publication outputs and impact measures.
   d. Align funder, institutional and charity reporting requirements with those of statutory reporting such as HESA returns and REF
   e. Ensure compliance with agreements to collect a minimum core dataset.

R3. Provide structures (common APIs, shared services or connectors) to support the exchange of research information, but not a central reporting system.

   a. Do not create a single national reporting system.
   b. Provide common APIs to source, not transformed, research information.
   c. Any technical solution for data exchange should be straightforward and have low integration costs.
   d. Provide a single point of deposit for research outputs.

R4. Increase the quality and timeliness of research information across the sector.

   a. Improve quality control of research information.
   b. Reduce human effort and increase automation in collection and processing of research information.
   c. Implement administrator workflows to reduce possibility of human error.
   d. Enable researchers to view and correct their own research information.
   e. Use of shared services for validation and quality control.
   f. Enable institutions to collect and validate research information prior to submission to funders.
   g. Enable ongoing reporting of research outputs to support ad hoc reporting by funders.

R5. Facilitate the flow of information between internal institutional systems and external systems (eg funder systems) in CERIF format.

   a. Integrate internal systems with CRIS to reduce re-keying and enable institutions to collate information for reporting.
   b. Enable bulk upload of data from CRIS systems to funder systems.

R6. Enable institutions to more effectively consume and reuse research information (eg for benchmarking and management information, portfolio management, collaboration, compliance monitoring, communications).
   a. Support data harvesting of data from multiple funder systems.
   b. Support for data harvesting from other institutions.
   c. Provide benchmarking tools.
   d. Provide ability to analyse data in different ways (eg according to department, collaborative network).
   e. Provide support for communications and marketing.

R7. Support benchmarking and portfolio analysis across research funders.
   a. Enable funders to harvest research information from other funder systems.
   b. Support benchmarking across funders.
   c. Support research portfolio analysis across funders.
   d. Support measure of long term impact of research.

R8. Provide appropriate data governance, transparency and security when collecting, sharing and reusing sensitive research information.
   a. Ensure compliance with data protection legislation.
   b. Protect the confidentiality of commercially sensitive data.
   c. Maintain trust of researchers in the use of the data.
   d. Provide retention policies to support long-term monitoring.
   e. Provide rigorous validation of data before release into the public domain.

Use cases

A number of use cases relating to the collection and reuse of research information were identified during phase 1. The following use cases for institutions were identified:

- Reporting to research funders
- REF reporting
- Statutory reporting to HESA (eg HE-BCI survey)
- Internal reporting, benchmarking and operational management (eg optimising facilities usage)
- Benchmarking against other institutions
- Portfolio analysis (internal) and collaboration (cross-institution)
- Submission of grant proposals to funders
- Researcher CV generation
- Strategic planning
- Compliance monitoring

The following use cases for research funders were identified:

- Publication of source data via common APIs
- Funder-funder benchmarking
- Portfolio analysis and strategic planning
- Reporting to government (regular and ad hoc)
- External communications

- Gathering of information from researchers
- Production of evidence of impact

These are described in more detail in Waddington *et al* (2013) and the UKRISS Feasibility Study (2013).

Technical review

As part of a more detailed study into the landscape, we conducted in-depth technical reviews of a number of key technologies in different aspects of research information management: CRIS systems, funder systems, business intelligence, information modelling and related Jisc-funded projects. In all we investigated 12 technologies/projects, more complete details of which can be found in Appendix D of the UKRISS Feasibility Study (2013).

- Elsevier Pure – institutional CRIS system
- Thomson Reuters Converis – institutional CRIS system
- BRUCE/SolrEyes – Jisc-funded research information project
- CERIF – research information data model
- InCites – research analytics service
- Je-S – research council grant submissions portal
- MICE – Jisc-funded research information project
- Readiness for REF (R4R) – Jisc-funded research information project
- Research Fish – commercial research outcomes system used by MRC and other research funders
- RMAS – Jisc-funded research information project
- ROS (Research Outcomes System) – in-house RCUK research outcomes system
- Symplectic Elements – institutional publications management system

The aim of this task was to survey the kinds of functionality that we find in well-developed widely used systems, and what potential implementation options we might have for any phase 2 development work. What we found was an extensive feature set which includes: authority lists for authors (on an institutional level), structured databases of citations and publications and flexible reporting and management information tools.

Commentary on feasibility study findings

The feasibility study in phase 1 uncovered a fragmented research information reporting landscape, with a lack of alignment in information requests made to institutions and researchers. Similar information is often collected in similar but different ways, placing a burden on researchers and research administrators in institutions. This diversity was also reflected in a lack of automation of key reporting processes.

Data quality was a major issue. In part this was due, in some cases, to low reporting compliance by researchers, resulting in incomplete data. In other cases, errors were caused by manual processing and lack of cross-referencing of data stored across different systems.

All of these issues limited the sharing and reuse of research information for purposes such as business intelligence, strategic planning and benchmarking.

Feasibility study recommendations

The UKRISS Feasibility Study (2013) made three recommendations for further work during phase 2 of the project and beyond:

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.

There are clear dependencies between the recommendations. Recommendations 2 and 3 depend on recommendation 1. Recommendation 3 can be implemented without recommendation 2, although issues such as data quality and compliance with the core profile schema would need to be addressed in other ways.

For each recommendation, the project developed a work plan combining different aspects of these objectives for investigation in phase 2. These plans are described in section 6 of the UKRISS Feasibility Study (2013).

Definitions: The feasibility study 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 that they could be mapped to a single reporting field.

Commentary on recommendations

An initial objective of UKRISS was to assess the feasibility of developing a national reporting service for research information reporting. 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:

- 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

- Much of the cost of integration with a national system would fall on institutions, and these costs were perceived as being potentially high

There was however support for common services amongst a wide range of stakeholders. In particular, funders were keen to have common API definitions for systems to support automated harvesting of reporting information.

Interoperability at a semantic level was seen as a prerequisite to developing national-level reporting services and benchmarking tools. Given the limited time and resources available, it was decided to focus primarily on recommendation 1.

However, given the abstract nature of recommendation 1 and the need to convince key stakeholders of the benefits and value of a core profile, supporting work on technical demonstrators and a business case was also seen as important for the second phase. Further, communication and consultation with key stakeholders was considered essential to obtain buy-in to the UKRISS approach.

3.2.2 Phase 2 objectives

A review of the feasibility study recommendations was conducted with Jisc in February 2013. The overall scope of the recommendations was very broad, and therefore an attempt was made to produce goals that were achievable within the lifetime of the project. Following the consultation with Jisc and discussions between the project partners, the following objectives were agreed.

1. Investigate opportunities for harmonisation between ROS, Research Fish and the HESA HE-BCI survey. Develop an aggregated profile and mappings to CERIF. Investigate opportunities for harmonising reporting fields and the extent to which this could be achieved. Where possible make harmonisation proposals and produce a core reporting profile. This would include:
    a. Reporting fields and vocabularies.
    b. CERIF mappings.
2. Develop tools and components to demonstrate the value of a core profile. These would include an investigation, and in some cases prototyping of:
    a. Validation of the profile, including schema compliance and data field validation.
    b. Visualisation of the core profile to enable research managers and application developers to more easily work with CERIF.
    c. Aggregation tools to demonstrate the potential for analysing information represented in a common format for business intelligence purposes.
    d. A Crosswalk Connector to demonstrate extraction and compilation of data into the core profile from institutional systems.
3. Understanding of the issues in producing a business case for harmonisation.

The aim of the technical work was to provide prototypes and demonstrators which could be developed further by the community rather than production software.

For objective 3, the initial aim was to produce a full business case. However, due to the uncertainties of the outcomes of the RCUK internal reporting systems review and the potentially large amount of effort required, it was decided instead to produce a more limited analysis of benefits, costs and risks. This work would support the scoping and development of a full business case at a later date.

3.2.4 Modelling methodology

In line with objective 1 in section 3.2.3, the goal of the modelling work package was the modelling of a core CERIF profile in support of harmonised reporting to RCUK and the HE-BCI survey. REF reporting was also included in this analysis. We aimed to develop an aggregated reporting profile (model) as a basis for harmonisation, as well as corresponding mappings to demonstrate its implementation in CERIF.

The investigation of inherent and potential degrees of harmonisation of the reporting landscape started from existing reporting objects and their describing fields within ROS and Research Fish. This included the controlled vocabularies associated with these fields, and comprised a bottom-up investigation at three levels:

- Level (a): reporting objects
- Level (b): describing fields

- Level (c): applicable vocabularies

The first results were collected bottom-up through level (a) where reporting objects subsumed the describing fields of level (b). Vocabularies applicable within fields at level (c) were separated out according to their quantity and heterogeneity. The resulting collection allowed for a "semantic similarity" degree calculation between ROS and Research Fish based on field comparison at level (b) and within reporting objects at level (a). In addition, an estimation of potential harmonisation degrees at level (b) anticipated the top-down modelling guided by a use case "Institution submits final report to funder" and the introduction of an upper reporting level at level (a). These are represented within the developed UKRISS Core Information Reporting Profile (figure 3.1).

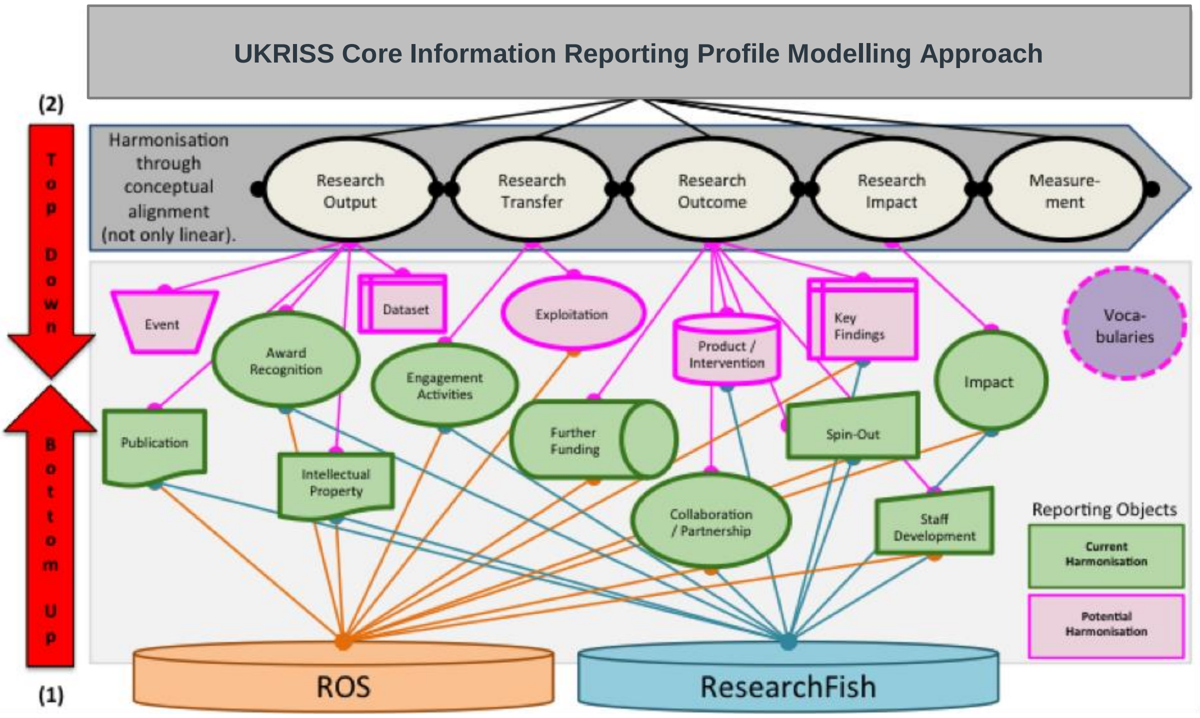


Figure 3.1: UKRISS Modelling Approach: Bottom-up (1) / Top-down (2)

The hybrid UKRISS modelling and harmonisation approach first steps bottom-up (1) ROS and Research Fish, and continues top-down (2) through the upper reporting level within the UKRISS profile and use case. Harmonisation at level (a) is indicated with colours, where green means current harmonisation and pink potential harmonisation of UKRISS reporting objects.

3.2.5 Technical demonstrators methodology – validation, visualisation and aggregation

To demonstrate the value of the UKRISS models, three areas of technical development were carried out, each of which relies heavily on the notion that the data they work with are represented in a structured, consistent, well-understood form. Without such consistent structuring, the following technical activities would be difficult or borderline impossible, or at least have very limited value.

All of the work carried out was limited to the UKRISS Research Outputs model, mostly due to limitations of time, although all the work here is extensible to the other model types. Additionally, all of the tools are proof-of-concept only, and while software has been made available open source, it is not in a stable or trivially reusable condition.

The three areas that were considered for the proof of concept work were:

Validation – the process of ensuring compliance with the model and the quality of the data represented.

Visualisation – the process of turning the model into a visual representation.

Aggregation – the process of bringing together and exploring many objects which are all expressed using the model.

3.2.6 Technical demonstrators methodology – Crosswalk Connector

As described above in the data validation, visualisation and aggregation section, the UKRISS project has not attempted to develop production-ready software as this was not in the project remit. However, it was considered important to demonstrate that the theoretical aspects of data structuring will be of benefit in a live environment, and this necessitated the development of proof-of-concept systems.

The intention was to demonstrate practically how institutions could automate the process of collating information required by external bodies. In the first instance, the focus was on the ROS and Research Fish returns, but the principle applies to many of the hundreds of data submissions based on data stored within institutions in an electronic format eg DLHE, HE-BCI, etc.

Working with a leading software supplier, Certus Technology Associates, the project team developed a "Crosswalk Connector". This is based on open source technology and could be made available to the HE community free of charge.

This Crosswalk Connector builds on the previous Jisc-funded development of RMAS and is designed to extract information from source systems within institutions, convert to a common data model (CERIF) and to deliver the structured information to a destination that may be accessed by funding bodies (eg research councils).

One aim of the Connector development was to minimise the barrier to entry for institutions wishing to install and use the Connector, and understanding the resource requirements for adoption was one of the required outputs from the proof-of-concept.

The Connector was deployed at the University of Exeter and used to interrogate live research systems such as Symplectic, transform this raw data into the structured CERIF data returns outlined in section 3.2.4, highlight any errors in the data, and to generate an xml output that could be passed directly to the funding bodies or alternatively passed on to additional data validation tools described in section 3.2.5. Further details are presented in section 3.3.3.

3.2.7 Business case methodology

Context

Implementation of a harmonised approach to research information reporting will require change to existing practice within the sector. For a harmonised approach to research reporting to be adopted within the sector, both the research councils and institutions as well as other funders and sector organisations need to appreciate the benefits and costs involved – the business case for change needs to be clear.

There are three broad options for adoption of harmonisation:

- A harmonised approach based on an integrated approach adopted by RCUK funders
- A harmonised approach managed by an open standards body on behalf of the UK research community
- The current status quo

While the feasibility study of UKRISS phase 1 identified the opportunities that harmonisation could afford, there were still a number of outstanding questions to be answered before undertaking a detailed business case.

- Whether the seven research councils will adopt one integrated solution for collecting data from institutions/researchers in the near future
- What risks are involved for organisations adopting harmonisation or not adopting; and how these risks differ for open standards based and non-standards based harmonisation; these risks will vary by type of organisation and size, with the risks and potential impact for non-RCUK funders (eg charities) and smaller institutions being particularly unclear at present
- The full extent of possible indirect benefits and how these might be measured to best convey their value
- What types of costs will be incurred in implementing harmonisation and how these will vary with organisation
- What the barriers to constructing an effective business case are and how these might be overcome

Aims and objectives

The aim of the preparatory business case work with UKRISS was therefore to explore the benefits and risks associated with implementing a harmonisation of research information reporting and data collection across key sections of the public research base, based on open standards and vocabularies. This would help address the current uncertainties and lay the groundwork for construction of an effective business case once RCUK's future plans regarding research reporting are known.

Specific objectives are:

- Describe the benefits (including negative benefits) for each of the stakeholders and how these might be measured
- Describe the risks for each of the stakeholders, their potential impact and how these might be mitigated
- Identify how the costs of implementing harmonisation might be measured
- Summarise the barriers to uptake for each of the stakeholders, with suggestions of how these might be addressed

Approach

As well as drawing on extant work such as the UKRISS phase 1 feasibility study, stakeholder analysis and data gathered from the recent UKRISS phase 2 workshops, we conducted a series of interviews with key stakeholders to explore potential benefits, risks and barriers. The preparatory business case work focused on three key stakeholders: the main non-commercial funders of research-RCUK funders and charities as well as institutions (both research intensive and less research intensive). Inclusion of charities was important as it allowed exploration of whether a RCUK integrated solution would also meet the needs of other funders. The number of new interviews (five in total) was limited due to availability of an appropriate representation of stakeholders within the timescales within which the work was carried out. The semi-structured interviews covered benefits and dis-benefits, risks, barriers and enablers, and costs. Participants were also asked to identify how benefits and costs might be measured within their organisations. To compensate for the limited stakeholder coverage in the interviews, an analysis of previous interviews undertaken as part of the feasibility study was undertaken to try to ensure all stakeholder groups were represented. The analysis consisted of identifying and categorising both direct and indirect benefits and dis-benefits. Benefits realisation mapping was then used to explore further the dependencies and risks to realising the projected benefits.

The outcomes and findings of the business case work-package are summarised in section 3.3.4.

3.2.8 Approach to dissemination and sustainability

Dissemination

Communication was seen as a key objective of UKRISS from the outset. As well as the technical goals of analysing and modelling reporting requirements and developing technical demonstrators, it was seen as critical to obtain buy-in from representatives of key stakeholders for the approach taken by the project. A dissemination plan was developed early in the project to describe the communications channels by which each of the key stakeholder groups would be addressed, as an internal report.

The UKRISS project blog (http://ukriss.cerch.kcl.ac.uk) was the primary medium for disseminating project progress, outputs and outcomes to a wider audience across the sector. The blog was updated on a regular basis, often weekly, and contained articles about project activities as well as providing a repository for formal and informal project outputs. The UKRISS Twitter feed (@ukriss_jisc) was used to publicise the blog posts and other project activities. Although we have no specific figures about the readership of the blog, we did receive a number of enquiries about the content from across the sector.

The primary means of direct engagement with stakeholders was through the UKRISS Steering Board, which included senior representatives of key stakeholders (see Appendix 1). The composition of the Steering Board was discussed and agreed with Jisc prior to the commencement of the project to ensure a balanced representation. The membership was also widened during the course of the project.

A final project workshop was held at the British Library in London on 1 November 2013. The workshop aimed both to present the project results to a cross-section of stakeholders across the sector as well as to gather additional feedback on the project results. Following the London workshop, a request was made to rerun the event in Scotland. This event was held at the University of Glasgow on 15 November 2013. This workshop also produced useful and relevant feedback.

Sustainability

The main engagements relating to sustainability are described below. Further details on specific activities and outcomes are described in section 6.3.

CASRAI

Consortia Advancing Standards in Research Administration Information (CASRAI) is a non-profit standards development organisation (see http://casrai.org). CASRAI is an international community of leading research funders and institutions collaborating to ensure interoperability of research information. A 12 month UK pilot funded by Jisc was run from July 2013 to June 2014. Representatives of key sector bodies such as RCUK funders, institutions and charities are participating. UKRISS was represented in both the Contributions and Open Access Task Group as well as in the Review Circle.

HEDIIP

The Higher Education Data and Information Improvement Programme (HEDIIP) has been established to enhance the arrangements for the collection, sharing and dissemination of data and information about the HE system. HEDIIP is leading a programme of changes to build a more coherent, responsive and less burdensome information landscape.

Although the primary focus of HEDIIP is on teaching, there are significant areas of overlap with the work of UKRISS in such areas as reporting related to research students as well as in vocabularies. UKRISS engaged with HEDIIP through a series of face-to-face meetings, which resulted in a contribution to the report: The HE Information Landscape: Creating and Managing a Data Model (2013).

Engagement with funders (RCUK, charities) and institutions

The primary direct engagement with funders and institutions was through the Steering Board and related contacts. The project carried out a wide scale consultation in phase 1, interviewing over 40 subjects, which resulted in collection of a wide range of inputs from funders, charities and institutions. In phase 2, engagement at a more detailed level has taken place. Although the Scottish and Welsh national bodies and funders were not interviewed in the phase 1 study, we were able to obtain feedback from the Scottish Funding Council at the UKRISS Glasgow workshop. There are also other significant research funders in the UK such as the European Commission and government departments, with whom we have not engaged directly but have an interest in the overall outcomes.

Partner institutions

Exeter led the work on building the Crosswalk Connector, which enables research information to be gathered from an institutional system, mapped to CERIF, and assembled into the appropriate form for reporting to funders. Exeter have an ongoing internal project to further develop the Crosswalk Connector beyond the end of UKRISS.

The partners King's College London (KCL) and Brunel University have provided sample data from their institutional systems for use in designing and testing the Crosswalk Connector. The KCL dataset was a full set of MRC reporting data for 2011/12. The Brunel dataset was a set of reporting information extracted from their Symplectic CRIS system.

At the end of the project and beyond the end of the funding period, KCL and Brunel are working on integrating internal demonstrators of the Crosswalk Connector, in order to evaluate its suitability for production use. This work is being supported by Certus, who were subcontracted to carry out technical work on UKRISS by Exeter. The validation and visualisation tools will be integrated with the Crosswalk Connector to demonstrate additional quality assurance capabilities.

euroCRIS

A presentation of UKRISS was given at the CRIS 2012 conference, and a submission has been made for CRIS 2014 to raise awareness of the UKRISS work amongst the wider international community. The UKRISS core profile does not require extensions to the CERIF standard as such. However, a suggestion has been forwarded to the CERIF TG (collocated to the November 2013 membership meeting in Porto), which received a lot of positive interest in particular by suppliers, with regard to application of "business" rules on top of CERIF objects in support of object-aggregation following use case driven requirements. However, the specific application of the standard is of wide international interest, as many other countries are facing similar harmonisation issues. The involvement of euroCRIS as a UKRISS partner will ensure the knowledge gained will be sustained and used to benefit the international research community.

3.3 What did you learn?

In this section we provide an overview of the findings from phase 2 and lessons learnt, based on the tasks described in section 3.2.

3.3.1 Core Information Reporting Profile

The UKRISS Core Information Reporting Profile (figure 3.2) is aimed at enabling harmonisation of ongoing UK research reporting activities. It is the result of a thorough investigation of current reporting systems and activities. The bottom-up analysis of ROS and Research Fish reporting objects uncovered a "semantic gap" between their object describing fields (see figure 3.1 in section 3.2.4) and revealed issues around developing a harmonised information reporting profile in a bottom-up manner. A review of related initiatives and projects suggested a more conceptual or top-down view to balance out these semantic differences, see the REF Impact Pilot Exercise (2010), DESCRIBE (2013), MICE (2011) and Duryea *et al* (2007).

A use case was adopted to motivate the UKRISS modelling activity: "Institution submits final report to funder". The use case guided the profile development conceptually (top-down) and resulted in the addition of an upper reporting level above the reporting objects in support of semantic clarity and consistency with the subsumed reporting objects' structure. This ensured that the usage of the reporting objects followed the requirements.

The introduction of the use case as well as the upper level model extension enabled a much more meaningful understanding and thus reuse of comparable elements or (sub-)objects, eg from REF reporting (eg "institution submits report to funder") and the HE-BCI survey (eg spin-off activity vs spin-out in UKRISS).

Feedback from RCUK has been taken into account by means of adjustments to the final model. It clarified in particular the model's usage with sub-typing of reporting objects according to upper level concepts. These turned out to have been biased by an institutional (data producer) lens in earlier versions following the bottom-up investigation from a submitter's view. The earlier assignments have finally been adjusted towards a more top-town perspective of a funder (data consumer) collecting reports from institutions, for whom the harmonisation profile and the reported data are intended and by whom the wider adoption of the harmonisation profile will hopefully be supported, to then guide implementation at institutions.

Figure 3.2 presents the UKRISS Core Information Reporting Profile including the upper level of reporting concepts Research Output, Research Transfer, Research Outcome, Research Impact and Measurement, under which the reporting objects are subsumed.

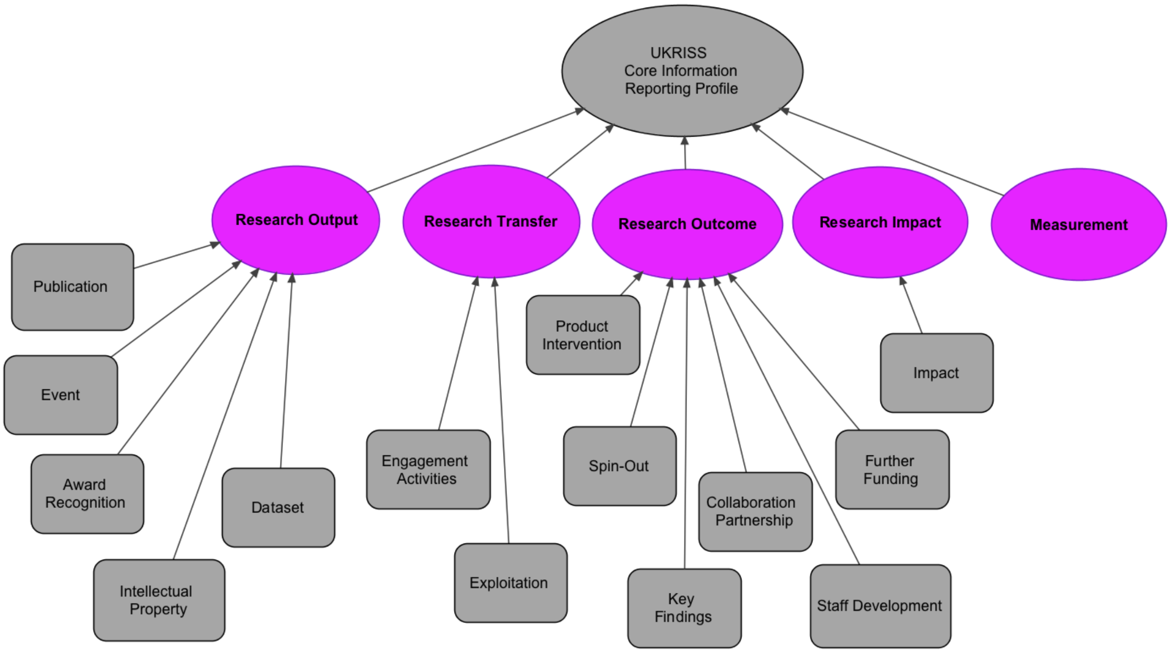


Figure 3.2: UKRISS Core Research Information Reporting Profile

Reporting records are instantiations of UKRISS reporting objects and their assignments to upper level reporting concepts do not necessarily follow a linear process or sequence (see also figure 3.1 in section 3.2.4). A clear or single assignment in practice may very much depend on a use case as we have also learned ourselves; eg an institution could consider "Staff Development" as a "Measurement", whereas a funder could see it as "Research Outcome". Different or multiple conceptual assignments may thus be required or even suggested in support of and depending on multiple involved stakeholder viewpoints.

The CERIF XML examples anticipate the funder's viewpoint reflected in figure 3.2. The definitions behind the upper reporting concepts are available in Appendix 4.

Harmonisation degrees

UKRISS defines harmonisation as "semantic similarity" between comparable entities, where entities can be objects, fields, or applicable vocabularies. Figure 3.3 reveals current harmonisation degrees (green) between ROS (blue) and Research Fish (red) reporting objects based on object describing field counts, and anticipates potential harmonisation degrees likewise through a UKRISS Core Information Reporting Profile (purple). Figure 3.4 illustrates harmonisation degrees from an anticipated implementation of the UKRISS Core Information Reporting Profile implying an introduction of changes that include harmonised controlled vocabularies to ensure consistent implementation of reporting records in compliance with CERIF.

Similarity examples at field and vocabulary level:

- **Fields within reporting output object "Publication":**
    - "Provide a short name/title for this output" *is similar to* "Series Title"

  - "Year of Publication" *is similar to* "Date Published" or
  - "Volume" *is similar to* "Volume Number"

- **Vocabularies within reporting output object "Publication":**
  - "Publication Type" *is similar to* "Type of Publication"
  - "International Audience" *is similar to* "What was the extent of geographical 'reach' of this activity (eg region, nation or international)"

A more detailed description of the UKRISS harmonisation analysis is available in Appendix 3.

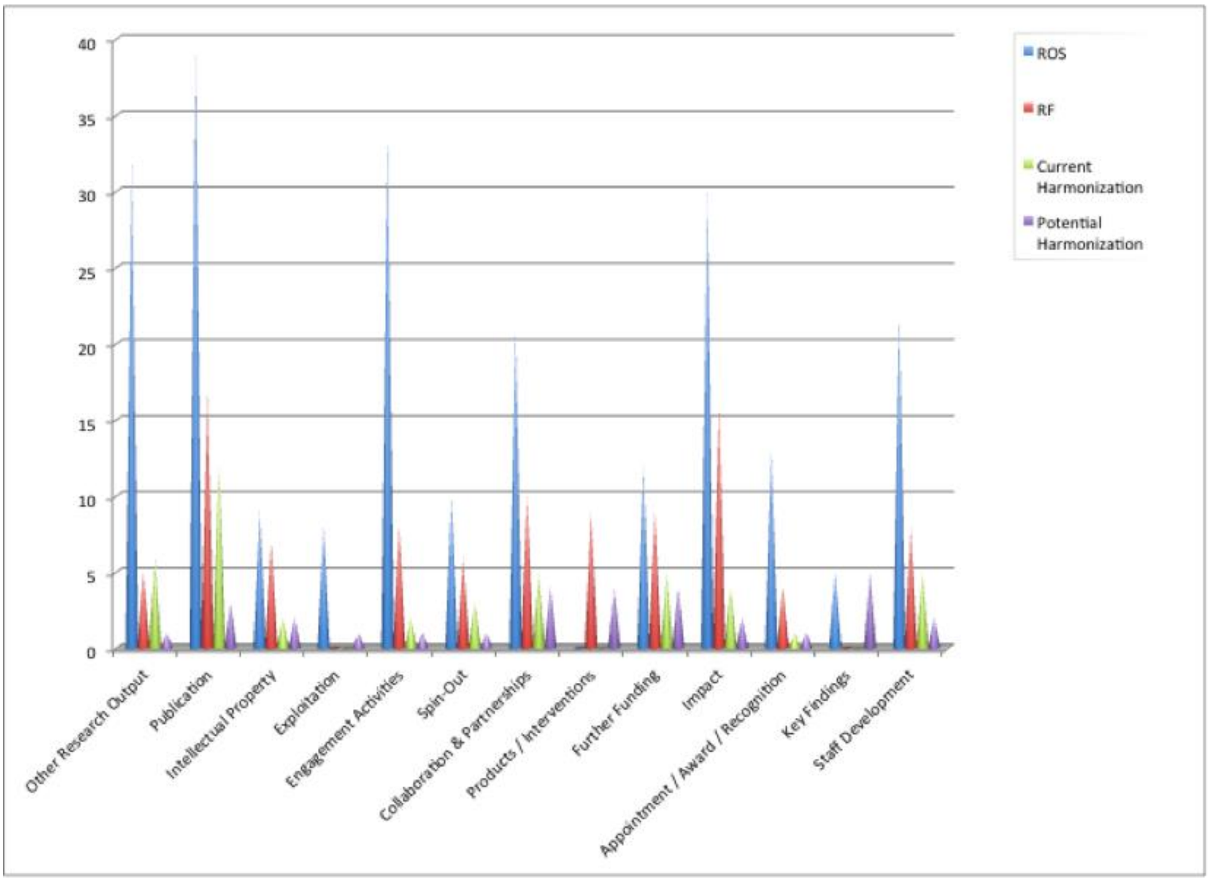


Figure 3.3: ROS/RF reporting object harmonisation degrees before UKRISS

In addition to introducing an upper reporting level UKRISS suggests a header as well as stakeholder and generic references with each reporting object (see Appendix 2). A potential harmonisation degree in line with the UKRISS profile implies a normalisation at object field levels following the CERIF structure, ie identifier-driven object aggregation and thus flagging of redundant string fields; federated identifier application and the introduction of vocabularies for relationships.

A harmonised UKRISS profile therefore flags "Other Research Output" as a reporting object (see figure 3.4 where it has no more fields). It introduces "Generic Fields" on top of each object and vocabularies to avoid extensive field application in compliance with CERIF.

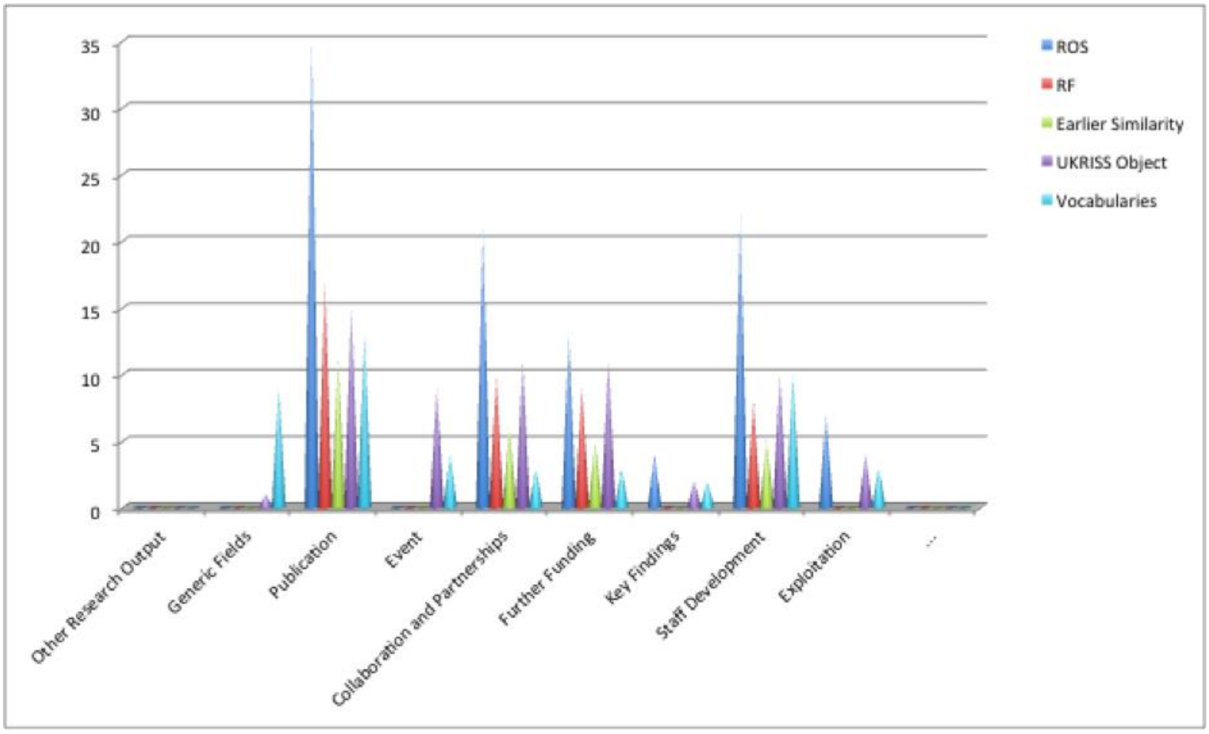


Figure 3.4: ROS/RF reporting object harmonisation degrees through UKRISS (not complete – a few examples only)

In figure 3.4 the earlier ROS/RF harmonisation degrees (green) from figure 3.3 are preserved. Figure 3.4 anticipates the degree of potential harmonisation through CERIF-compliant restructured UKRISS objects (purple). The Excel spreadsheet (Appendix 4) provides more details behind the charts and a comprehensive overview of the UKRISS Profile elements and definitions.

Figures 3.5 and 3.6 inform about total harmonisation degrees at field and vocabulary level within ROS and RF.

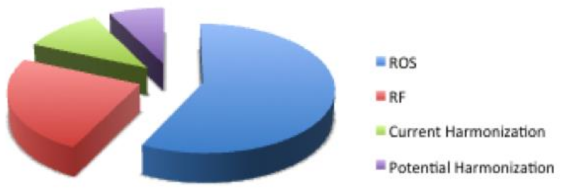


Figure 3.5: RS/RF total harmonisation degrees at reporting object's field level.

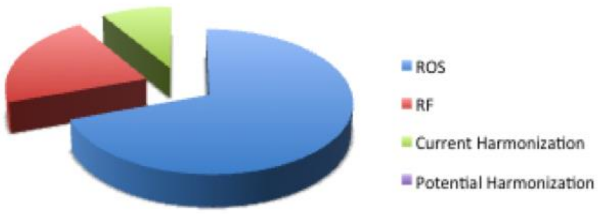


Figure 3.6: RS/RF total harmonisation degrees at reporting object's vocabulary labels level.

Table 3.4 provides the counts behind figures 3.5 and 3.6. The initial harmonisation analysis between ROS and RF counted 45 current fields and 31 potentially harmonisable fields from the 235 ROS and 99 RF fields. Only 5 of the 39 ROS and 12 RF vocabularies were considered harmonisable according to the initial (preliminary) investigation at field and vocabulary label level – not vocabulary term level (which needed much deeper investigation and was hence not countable and thus remained empty in table 3.4). The potential harmonisation degrees at field level (left column) were counted from fields inherently found harmonisable. These included unique identifiers as well as potential vocabulary extensions.

Total Field Counts		**Total Vocabulary Label Counts**	
ROS	235	ROS	39
RF	99	RF	12
Current Harmonisation	45	Current Harmonisation	5
Potential Harmonisation	31	Potential Harmonisation	–

Table 3.4: ROS / RF field and vocabulary counts

Recommendations for implementation

UKRISS investigated inherent and potential degrees of harmonisation within the mentioned UK reporting systems and suggests a conceptual gradual approach towards harmonisation through structural changes and extensions following and supporting technological, standardisation and identifier trends.

These changes imply the introduction of upper level reporting concepts, generic elements to all reporting objects, and a normalisation of objects through the replacement of some fields through relationships and applications of unique identifiers, as well as the introduction of reusable controlled vocabularies, in line with CERIF. Explicit change counts as visualised in figure 3.4 are available from within the Excel sheet (Appendix 4).

In addition to extracted (bottom-up) reporting objects we recommend Dataset and Event to be added as new reporting objects. We furthermore suggest the impact reporting object should be approached with more care especially towards reflecting timelines that allow recording and time-tracking beyond project boundaries.

From feedback received during project presentations and engagement activities we recommend Equipment and Education as important future reporting objects, thus bridging the current gap with Higher Education reporting and Research Infrastructures support as such.

Before any implementation we recommend a clear harmonisation of requirements across the sector involving a maximum number of key stakeholders to guide the priorities and the approach towards implementation of the proposed extensions and the rules to apply.

3.3.2 Validation, visualisation and aggregation demonstrators

In this section we will show the value of the validation, visualisation and aggregation demonstrators described in section 3.2.5. Each of these technical aspects aims to make the case that the UKRISS models are a viable and sensible step forward for the research information community. More detail on the software and approaches we have used is available in Appendix 5.

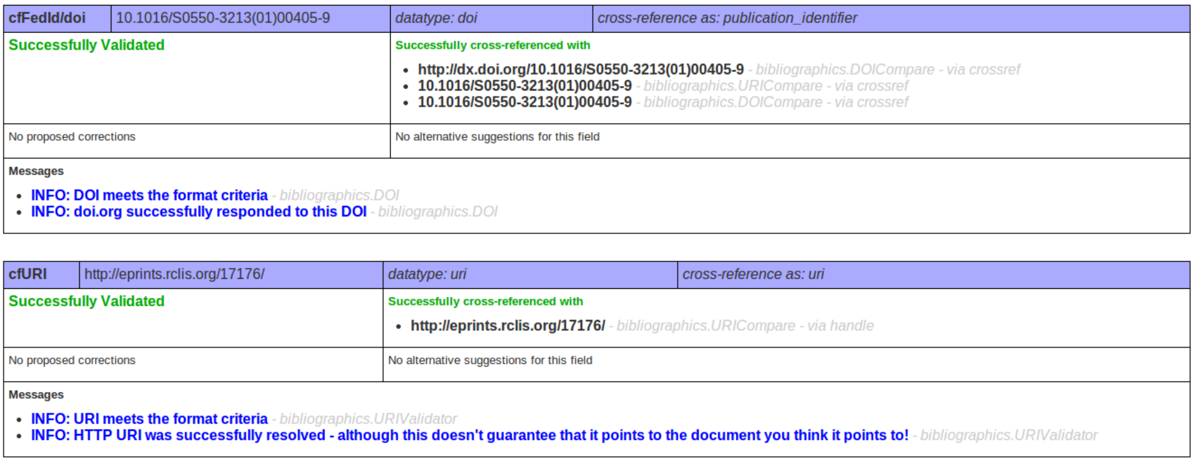
Validation

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 (eg knowing when a field is supposed to be a date, and even in what format that date should be represented), means that generic validation software can be developed, which can then be used by any organisation that works with research information.

The benefits of validation are several fold:

1. It can catch errors with metadata at the point that it is created or stored, preventing it from propagating incorrectly.
2. It can improve your metadata quality at source.
3. It can help ensure consistency of metadata across the whole community, which in turn will be of value in other contexts, such as reporting and analytics.

The validation software that we developed during the course of this project is proof-of-concept but quite powerful. It can validate individual elements of the model (such as checking that an ISSN is an ISSN), but it can also look-up data in external data-sources (such as Entrez[1]) and cross-reference that with the document being validated, which helps make metadata consistent across the whole community.



cfFedId/doi	10.1016/S0550-3213(01)00405-9	*datatype: doi*	*cross-reference as: publication_identifier*
Successfully Validated		**Successfully cross-referenced with** • **http://dx.doi.org/10.1016/S0550-3213(01)00405-9** - *bibliographics.DOICompare - via crossref* • **10.1016/S0550-3213(01)00405-9** - *bibliographics.URICompare - via crossref* • **10.1016/S0550-3213(01)00405-9** - *bibliographics.DOICompare - via crossref*	
No proposed corrections		No alternative suggestions for this field	
Messages • **INFO: DOI meets the format criteria** - *bibliographics.DOI* • **INFO: doi.org successfully responded to this DOI** - *bibliographics.DOI*			

cfURI	http://eprints.rclis.org/17176/	*datatype: uri*	*cross-reference as: uri*
Successfully Validated		**Successfully cross-referenced with** • **http://eprints.rclis.org/17176/** - *bibliographics.URICompare - via handle*	
No proposed corrections		No alternative suggestions for this field	
Messages • **INFO: URI meets the format criteria** - *bibliographics.URIValidator* • **INFO: HTTP URI was successfully resolved - although this doesn't guarantee that it points to the document you think it points to!** - *bibliographics.URIValidator*			

Figure 3.7: Example output of the metatool validation process

[1] http://en.wikipedia.org/wiki/Entrez

Visualisation

This is the process of converting documents formatted using the UKRISS model into graphical representations. We were interested in being able to visualise both documentation of the UKRISS models and any expression of data using them (eg a publication record). Having data in a well-structured format means that we can have generic visualisation software that can be used by anyone working with CERIF data.

The benefits of visualisation are several fold:

1. To document the UKRISS outputs.
2. To enable data producers to visualise their own content.
3. In both (1) and (2) to allow end-users to explore the data in a more user-friendly manner.

The visualisation software that we produced during this project will take any CERIF XML (in the 1.6.2 version) and convert it into a collapsible-tree visualisation where the end-user can click to expand or contract nodes, and to see the relationships between objects and their classes or other objects in the model. While it is still relatively basic, it shows the potential for a more comprehensive CERIF dashboard which could be embedded into research information systems to aid comprehension of the records.

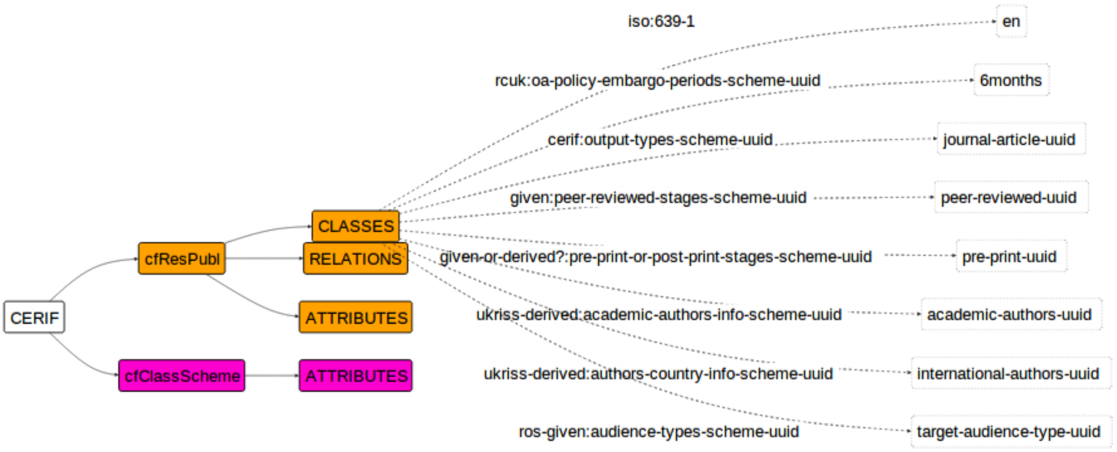

iso:639-1
en
rcuk:oa-policy-embargo-periods-scheme-uuid
6months
cerif:output-types-scheme-uuid
journal-article-uuid
given:peer-reviewed-stages-scheme-uuid
peer-reviewed-uuid
CLASSES
cfResPubl
RELATIONS
given-or-derived?:pre-print-or-post-print-stages-scheme-uuid
pre-print-uuid
CERIF
ATTRIBUTES
ukriss-derived:academic-authors-info-scheme-uuid
academic-authors-uuid
cfClassScheme
ATTRIBUTES
ukriss-derived:authors-country-info-scheme-uuid
international-authors-uuid
ros-given:audience-types-scheme-uuid
target-audience-type-uuid

Figure 3.8: Portion of the visualisation for a research output object

Aggregation

This the process of bringing together and exploring many objects which are all expressed using the model. What aggregation is not is the pooling together of differently formatted objects in the same space – with a heterogeneous collection like that, analytics, reporting or interesting views on the data are not possible because of a lack of common understanding. Therefore, a well-structured and well-understood data model is a requirement before those activities can take place.

The benefits of aggregation are almost unbounded:

1. You can do analytics on your own organisation's data objects for internal reporting.
2. You could look at cross-organisational data to compare and contrast organisations with each other, or report within subject domains.
3. Different funders with different requirements for reporting could extract the aspects of the data they are interested in from a large institutional aggregation.
4. Users of the aggregation could custom build their own reports and analytics for a wide range of diverse use cases that we haven't even thought of yet.

To demonstrate the power of aggregation, and the diversity of the kinds of questions that could be asked of a useful aggregation, we chose an unusual use case which was collected by the G4HE[2] project (a UKRISS sister project). We reworded it slightly to reflect usage of the publications model which was the focus of much of the technology work:

"As a postgraduate applicant I want to find out which institutions publish in my subject area because it will help me decide where to apply"

We then went on to produce a tool whereby a graduate could upload, for example, their masters thesis, and we would text mine it and then cross-search publications data from multiple institutions to rank them by applicability to that student's interests.

Conclusions and further work

These software outputs demonstrate the range and power of the tools that could be developed if harmonisation of the models of research information interchange can be achieved. Beyond the benefits of improved exchange/reporting of information, we can do a better job of ensuring that information is correct (through validation), present it to end-users in easy-to-understand ways (through visualisation) and carry out advanced data analytics giving funders and institutions the opportunity for key business intelligence (through aggregation).

As the software developed during the project is proof-of-concept only, there is significant additional work required to bring any of them to fruition. Additionally, of course, they are mostly of value if the data fed into them have been harmonised. There is an opportunity to use the software as agents for change, though, by demonstrating the value of harmonisation.

The main focus of further work around this should be:

1. Further development of validation tools – even without harmonisation, these tools can have significant impact in the community, and be able to make data easier to interchange even in the absence of the UKRISS models.
2. Development of aggregation user stories and analytical dashboards (along the lines of G4HE) to answer questions valuable to the community (and in particular funders and HEIs).

3.3.3 Crosswalk Connector

Overview

[2] http://g4he.wordpress.com

This section describes the approach to the development of a proof-of-concept for a Connector, introduced in section 3.2.6, that could be used by any institution (or funding body) within the HEI sector.

The University of Exeter lead the development of an open source connector 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, eg funding body. This methodology is shown schematically in figure 3.9.

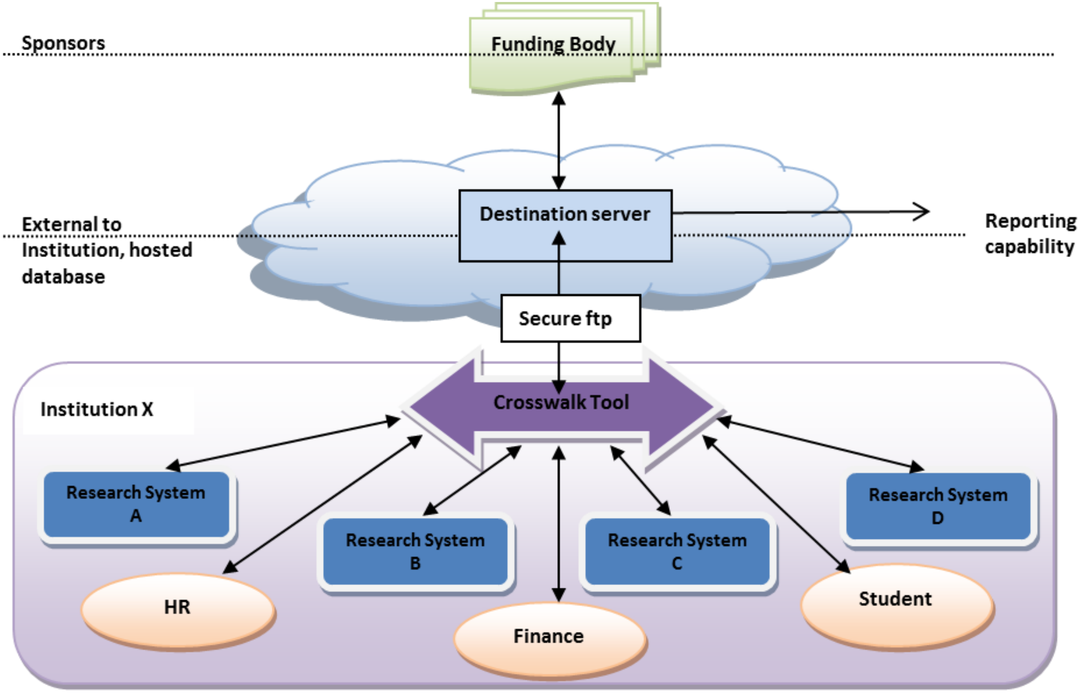


Figure 3.9: Schematic of the Crosswalk Connector interfaces

Even within the limited use-case of the ROS and Research Fish returns, the issues addressed by such a connector are multiple:

- Many forms of information flow between the institution and funders, particularly in support of research reporting, are based on manual processes
- Responsibility for completion of many reporting streams is split between Principal Investigators and central research support administrative staff, with little visibility of data entered
- Much of the required data resides across a range of internal systems, but each typically storing the data in different database structures and with different naming conventions for data fields
- Manual processes inevitably lead to
  - Inconsistencies in data quality
  - Ineffective reporting at an institutional level and also difficulty in drawing meaningful comparisons between institutions
- Significant potential for productivity savings and improved intelligence on research activity through automation and standardisation of data structures

Clearly, these issues also apply to other data returns and information exchanges based on manual extraction and entering of data, beyond those required for ROS and Research Fish.

The guiding principles for the connector 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. This enables the connector to be valuable not only to research intensive institutions, but also smaller organisations that may not have the same degree of technical capability. Using open-source technology also leads to a degree of future-proofing in that the community may wish to share deployment experiences and configurations as part of the natural progression of the product.

The proof of concept was installed and operated at Exeter, with work to demonstrate operations in other institutions (King's and Brunel) extending beyond the lifetime of the UKRISS project. For this proof-of-concept, there was currently no intention to interface directly to a funder's system, rather the information will be securely transferred to a known destination server, from which the funder could access the data. Clearly, interfacing directly into a funder's system would be a next logical development step, should the sector wish to proceed towards a production solution.

A more detailed technical description of the connector is provided in Appendix 6.

Conclusions and further work

We have shown how such a connector may provide a rapid and low cost means of automating data returns, assuming the source data are held electronically with an institution. The solution is sufficiently flexible to accommodate a range of data types and is therefore compatible with a wide range of existing systems (and spreadsheets).

One of the principal benefits of aggregating data from multiple sources derives from how effective this process is at highlighting errors and inconsistencies in the source data. Significant improvements in source data quality arise from implementation of this type of approach.

Interestingly, Brunel University and Exeter both use Symplectic for publications management. However, during the course of these trials it has become apparent how different the data formats are between these institutions, resulting from slightly different system configurations and use patterns. This highlights the power of transforming data into a Common Data Model and standardising on a format such as CERIF XML – without this approach comparisons between data, even from apparently similar systems, are extremely difficult.

The next step would be to extend the trials of the connector within the partner institutions and to include a wider variety of data sources and output content. Discussions are ongoing within King's College for the deployment of the connector for purposes beyond the UKRISS remit. Following more extensive trials, production versions of the software for wider distribution could be developed.

3.3.4 Preparatory business case

In this section we summarise the key findings of the preparatory business case work package. Full details are provided in Appendix 7.

Benefits

The main benefits of harmonisation for the key stakeholders examined are listed in table 3.5.

Benefit	RCUK funders	Charities	Institutions
More efficient collection of better quality data on research	y	y	Y
Better able to demonstrate performance of research councils' investments in research	y	–	–
Benchmarking and business intelligence analysis to help improve performance	y	y	Y
Better able to disseminate research outcomes	y	y	Y
Remove the need for niche information specialists	n	n	Y
Improve the organisation's research profile	y	y	Y

Table 3.5 Summary of the main benefits for the key stakeholders explored in the preparatory business case work package

For all the stakeholders, the benefits of harmonisation are likely to outweigh the dis-benefits. All the stakeholders who would use harmonisation (as opposed to developing systems) are likely to see direct benefits of efficiency and improved quality and coverage of data. This will enable them to benchmark their performance and explore how to improve research outcomes. Of particular interest is that the harmonisation model could potentially help improve impact reporting, especially where collaboration is involved – something which has proved difficult to do effectively to date. Although harmonisation potentially affords process efficiencies and new analysis capabilities, to make the most of these organisations will need to further develop their information systems and legacy data.

While considerable potential benefits have been identified, valuing these is more problematic. The work undertaken as part of the business case for implementing CERIF wrapper in Bolton (2010) provided some tentative details of efficiency savings; however, little work appears to have been carried out in institutions to estimate the value of benchmarking or business intelligence. Economic shadow pricing techniques could be used to find indicative pricing when developing the full business case. See Appendix 7 for more details.

Dis-benefits

As table 3.6 illustrates, there are also some potential dis-benefits associated with harmonisation.

Dis-benefit	RCUK funders	Charities	Institutions
Ability to reuse data on research reporting inappropriately	y	y	y
More difficult to report accurately on individual research council activities and impact where cross-cutting investment is used	y	–	–
Potentially prohibitive cost for adopting harmonisation	–	y	y – small/less research intensive
Widen the gap between larger/research intensive institutions and			y – small/less

smaller/less research intensive institutions			research intensive
Loss of granularity in reporting	–	y	–
Loss in flexibility to change reporting requirements to adapt to changing need	y	y	y

Table 3.6: Summary of the main dis-benefits for the key stakeholders explored in the preparatory business case work package

Some concern was expressed that data on research could then be reused inappropriately. For example, there was a concern that by harmonising reporting on research, commercial confidential or private data could be exposed. Additionally, there was a worry that the collected data could be reused out of context. For example, the provenance or source of some data may be lost in the reporting process and so interim findings could be reported as final verified results.

Adoption of harmonisation based on RCUK needs could also potentially mean that more niche data used by individual charities would not be captured.

The potential to misuse and loss of granularity could be largely addressed by good information policies and participation in development of harmonisation specifications.

A more serious dis-benefit was the potential that harmonisation might widen the gap between research intensive and less intensive institutions. Similarly, charities could be disadvantaged if costs prevented them adopting harmonisation. Indeed, the cost of implementation was a dis-benefit identified across the stakeholder groups.

Costings

The types of costs involved in implementing harmonisation are reasonably consistent across RCUK funders, charity funders and institutions. These would typically include: software; implementation of new data models or mapping; training; maintenance; and staff time. For small institutions and charities the costs are a significant overhead and could limit uptake. For charities that use services such as Research Fish, it would be expected that any implementation costs would be borne by the service provider, perhaps giving rise to an increased service charge for accessing the "harmonised" system.

There will be additional costs to realise wider indirect benefits such as improving performance through business intelligence – eg costs associated with new hardware; refactoring of legacy data and re-engineering processes; and creation and maintenance of registries.

While the business case for implementing CERIF wrapper in Bolton (2010) provided some tentative costing for implementing the wrapper, interviewees and workshop participants all felt that the costs were unrealistically low. More work is required to develop a more realistic estimate of the cost of implementing harmonisation. Ideally, this should be undertaken in a partnership with institutions and funders. Without their cooperation, it will be difficult to establish realistic costings.

Risks

The main risk associated with implementing harmonisation was that the harmonisation model may not meet the reporting needs of all stakeholders. For example, for research councils it may make it more difficult to scope reports to accurately report on an individual council's investment in specific cross-cutting investments such as the funding of shared research facilities. Similarly, charities may not have the same granularity of detail in researcher profiles. Poor decisions or negative impact due to inappropriate analysis were also seen as a risk; however, this could be mitigated through appropriate information policies. There is also a risk that some anticipated benefits such as business intelligence will not be realised as they will require additional investment in institutional information systems or expansion of the core harmonisation model.

There is a significant risk that the cost of implementing harmonisation is prohibitive.

The main risks associated with non-implementation of harmonisation were less efficient research reporting and analysis systems. For RCUK funders there is also a risk that they would be less able to compete for government investment in research, while for charities the risk is that they would be less able to attract leading researchers.

Relatedly, there is also a significant risk that the core harmonisation model may not be sufficiently extensive to deliver anticipated benefits. For RCUK funders and indeed the government and wider economy, this risk relates to not being able to fully and effectively report on research impact. For institutions, there is a risk that they may not be able compete as effectively as they lack relevant business intelligence data.

In general, stakeholders felt that the risks of not implementing harmonisation outweigh any risks associated with implementing harmonisation.

Risks do not vary by size of funder; rather the likelihood of the risk becoming an issue does. It is different for institutions, with there being a risk of a widening research gap between large/research intensive institutions and small/less research intensive ones.

There is also a significant risk that a partially harmonised model developed by the research councils is adopted, resulting potentially in a “one supplier” solution, which does not meet the reporting and business intelligence needs of the wider HE research community. While such a solution might lead to a high level of uptake as it is required for reporting on research council funded research, without appropriate governance, other funders could find it difficult to adopt or have no voice in submitting changes that would make it suitable for their requirements.

Finally, there is also a risk that lack of sector-wide leadership on harmonisation may mean that a non-optimal approach to harmonisation emerges as well as uptake being limited due to lack of support and benefits realisation management.

Barriers and enablers

Cost of implementation and lack of a clear driver for implementation were viewed by far as the biggest barriers to uptake of implementation. Harmonisation of research reporting across all the research councils was viewed as the biggest enabler followed by development of CRIS systems that support the harmonised models and dictionaries. However, as the research councils need any major new investments signed off at a high level in government, an effective business case linked to improving research impact is key. Adoption by the research councils of their internal system rather than a UK wide harmonisation model that supports the requirements of all the stakeholders in research reporting could potentially act as a barrier to the full realisation of benefits across UK HE. A sector-wide leadership role which seeks to drive an optimal approach to harmonisation (cost-effective while meeting the whole sector needs) as well as provision of support information and benefits realisation management would significantly aid both effectiveness and uptake.

Summary and next steps

Research reporting is a key part of the research ecosystem as it helps manage the return on investment in research as well as disseminating research outputs. Many stakeholders are involved including: the government as the funder of research councils, as well as the councils themselves and other funders; research organisations and individuals; others such as vendors of research reporting systems; and business and society as the ultimate benefactors of the research. The business case for harmonisation of research reporting, therefore, needs to be situated within this complex innovation ecosystem, illustrating the business case for each of the key stakeholders as well as how it will contribute to increasing economic and social benefits to business and society.

The business case should consider three options:

- A harmonised approach based on an integrated approach adopted by RCUK funders
- A harmonised approach managed by an open standards body on behalf of the UK research community
- The current status quo

The research councils are currently investigating ways to increase harmonisation in the reporting of research outcomes across the research councils. Once their way forward has been announced, a more accurate business case for adoption of a harmonised approach to research reporting across the UK HE sector could be explored.

The business case should need to not only compare the value of the benefits against the costs of implementation for the key stakeholders involved, but it should also need to address the dis-benefits as well as identifying the enabler that need to be put in place if harmonisation is to be adopted and benefit the whole of UK HE.

Many of the costs and direct benefits can be directly priced. Economic shadow pricing techniques should be used where required to price the direct benefits. This will enable the cost/benefit ratio for the various options to be calculated for each of the key stakeholders. Measures of efficiency and effectiveness can also be explored.

The potential economic impact of the risks identified also needs to be considered. The risk analysis provides a starting point, and the direct dis-benefits can be shadow priced.

Development of such a business case would require the cooperation of key stakeholders – in particular representative research councils, research intensive and non-research intensive institutions, charities and CRIS suppliers. Without reliable data from these stakeholders it is highly unlikely that a sufficiently detailed business case could be developed. The business case therefore should be developed in partnership with representatives from each key stakeholder group.

3.3.5 Stakeholder engagement

CASRAI engagement

CASRAI was seen as the main destination for standardising both the core profile as well as the associated vocabularies developed in the project. A preliminary presentation of the results of UKRISS was given to the CASRAI UK Contributions and Open Access task group in November 2013.

The final version of the UKRISS core profile and associated CERIF mappings has been submitted for consideration by the task group in January 2014.

HEDIIP engagement

At the beginning of the engagement with HEDIIP, it was our hope that a parallel activity could be carried out to map and align reporting fields and vocabularies used in teaching to the research information core profile. It soon became apparent that data structures and information exchanges in the teaching area are far more complex than in research, with a much greater number of external bodies requesting information from institutions. The complexity of the teaching area implied the level of harmonisation that we were aiming for in UKRISS would not be achievable for teaching information, at least within the timescales of the project.

Two useful outcomes that were achieved in the discussions with HEDIIP were to identify the potential of CERIF for modelling teaching information, and to identify areas of overlap between teaching and research, such as reporting on research students.

The primary focus of the initial work carried out by HEDIIP was on identifying basic entities, such as producing definitions of "student" and "course". In some cases, there may be overlap between the UKRISS vocabularies and those under consideration by HEDIIP. In addition, a basic list of terms for the teaching area is maintained by the QAA (www.qaa.ac.uk).

Engagement with funders (RCUK, charities) and institutions

The UKRISS report will initially be reviewed by the Steering Board, which contains representatives of the key stakeholders. The RCUK funders are well represented on the Steering Board, and are the primary recipients of the core profile. The UKRISS core profile has already been discussed in detail with technical staff within RCUK, so we anticipate a close alignment with their requirements for harmonisation. Throughout the course of the project, there has been close engagement with RCUK staff at a technical level from a cross-section of the seven RCUK councils. In particular, we have ensured that input has been gathered from the councils using ROS and Research Fish.

Institutions were represented both on the Steering Board (including the ARMA ex-chair and UCISA-CISG chair), as well as through the institutional project partners (Brunel, Exeter and King's College London). Institutions are also strongly represented in the euroCRIS community. Through the project blog, two project workshops, published papers and presentations at sector-wide events, the aim has been to reach a wider audience.

Charities were represented on the Steering Board by Wellcome and AMRC (Association of Medical Research Charities). In phase 1 of the project, we interviewed representatives of a number of different charities to obtain a balanced view of their interests and requirements.

Partner institutions

The universities Brunel, Exeter and King's College London have provided subjects for requirements interviews, test data and support for prototyping and demonstrators. The institutions have different research profiles and so were able to provide diversity, both in terms of requirements, data sourcing and technical infrastructures. Work on prototyping of the Crosswalk Connector is planned to continue beyond the end of the UKRISS project.

euroCRIS

euroCRIS is a partner in UKRISS, so has already taken a leading role in disseminating the results of UKRISS as well as providing technical inputs on CERIF, and the latest developments from within the international CRIS community.

A poster paper was submitted to the CRIS2012 conference, and a full paper is planned for the CRIS2014 conference. Additionally UKRISS was presented at the euroCRIS partner meeting in Porto in November 2013.

3.4 Immediate impact

This section describes the immediate impact the UKRISS project will have on stakeholders compared to the situation before the project began. Since the project was comprised of two distinct phases, we consider the impact of these separately.

3.4.1 Phase 1 impact

Phase 1 of UKRISS produced an independent view of research information reporting across the HE sector based on a wide and representative consultation. The report was welcomed by key stakeholders as making an important contribution to understanding the issues of research reporting harmonisation. In particular, many of the key stakeholder organisations are represented on the UKRISS Steering Board. Although previous projects and reports have highlighted some of the issues around research reporting harmonisation, UKRISS contributed by producing recommendations based on a rigorous analysis of a large corpus of evidence gathered from across the sector.

During 2013, the councils in RCUK began their own internal review of current reporting systems as well as the opportunities for aligning their reporting requirements. Specific goals are to review business needs for reporting across the RCUK councils and the future needs for research outcomes systems. Although there are many factors that have contributed to this review, not least pressure from institutions, we believe that UKRISS has contributed significantly to this activity. Based on close engagement between RCUK and UKRISS project staff, the UKRISS work is informing and contributing to this work and we anticipate the final outcomes will be closely aligned. The results of the RCUK review are unlikely to be known before the end of the UKRISS project.

3.4.2 Impact at end of project

We have identified the following areas of impact at the end of UKRISS:

- The RCUK councils have initiated an internal review in parallel to the work in UKRISS, a part of which is the development of a common data model across all the councils. The UKRISS profile and associated CERIF mappings are likely to be a strong candidate for adoption in such a model

- The findings of the project have been presented to the CASRAI Reporting and Open Access Task Group and a formal standards submission will be made at the end of the project. The work on standardisation of such a profile is likely to extend beyond the end of UKRISS. The CASRAI UK pilot currently runs until June 2014, when we expect the final results to be announced

- The interaction with HEDIIP has identified areas of overlap between the information models in teaching and research. As a result of the UKRISS interaction, HEDIIP is investigating the use of CERIF for the modelling of teaching information. The vocabularies used in the UKRISS core profile will also be made available to HEDIIP

- The Crosswalk Connector, developed within the UKRISS project, has been developed and piloted at the University of Exeter. Pilots are planned at both King's College London and Brunel University. These pilots will also integrate the validation and aggregation tools. Work on these is planned to continue beyond the end of the UKRISS funding

- The UKRISS results have been presented at international conferences in Canada and Germany. Many countries are looking to the UK to provide a lead in the application of CERIF, and some such as Germany are launching their own harmonisation projects. The UKRISS approach and core profile have already attracted considerable interest. Reuse of the UKRISS principles and mappings would result in increased interoperability and information exchange across international borders

3.5 Future impact

There are four main potential future impacts of the UKRISS work:

1. Adoption of the UKRISS core profile by RCUK. At the time of writing the report, the councils are considering the adoption of a common data model. This would have an impact on the systems used by the funders themselves, simplify the reporting requirements for institutions and make the data more reusable by the wider community through increases in interoperability and data quality.

2. Adoption of the UKRISS Crosswalk Connector and technical tools (validation, visualisation and aggregation). Such tools would primarily be of benefit to larger institutions, which need to aggregate reporting information from multiple internal systems. The impact is dependent on the implementation of the core profile. The potential impact of this is most easily assessed by the pilot deployments at UKRISS partner institutions.

There is more indirect impact though. The UKRISS project outcomes form a key part of supporting the sharing and reuse of research information through providing a common information model.

3. Gateway to Research (GtR) is currently dependent on mapping information fields gathered from ROS and Research Fish. Two problems encountered are that the definitions of the information fields and vocabularies are often different meaning that the data cannot be easily aggregated and compared, and that the data are of low quality. An expected impact of adoption of the UKRISS core profile is that a greater number of fields will be directly comparable. Further, through use of the validation tools, the data quality can be significantly improved. As a consequence, data from GtR will be more useful for both the institutions and industry. In particular, the tools being developed by the partner project G4HE will have much greater impact and applicability as a result of more complete source data being made available.

4. There is a likely impact on non-RCUK funders through the adoption of the UKRISS core profile by RCUK. Research council grants form a significant proportion of the funding to research-intensive institutions, typically between 30 and 50%. The adoption of the UKRISS core profile would result in pressure on other funders such as charities to align their reporting requirements as far as possible with the RCUK profile. Hence we would anticipate a strong trend towards harmonisation outside of the original remit of UKRISS.

4. Conclusions

This section summarises the main conclusions of the UKRISS project classified into the categories specific conclusions, conclusions relevant to the wider community and conclusions relevant to Jisc.

4.1 Specific conclusions

1. The project produced a set of recommendations for harmonisation of research information, based on a large sector-wide consultation, demonstrations of the potential benefits and an understanding of the business cases for change, from national and organisational perspectives.

2. The project has defined a conceptual framework for aligning reporting fields across the RCUK outcomes systems and the HE-BCI survey based on a high-level classification of output types (Research Outputs, Research Transfer, Research Outcomes, Research Impact and Measurements). This improves the understanding and interpretation of existing fields and provides a basis for further work on harmonisation.

3. The project has carried out a detailed analysis and comparison of the reporting fields collected by the ROS and Research Fish systems and the HE-BCI survey, and determined the extent to which harmonisation can be achieved within the existing business requirements.

4. The project proposed a harmonised core reporting profile encompassing the RCUK reporting and HE-BCI, including alignment of reporting field definitions, CERIF mappings and vocabularies. Specifically we have:
   a. Mapped all existing reporting fields to CERIF, and defined common vocabularies where possible.
   b. Made proposals for alignment between similar reporting fields.
   c. Indicated the extent to which harmonisation can be achieved within the existing business requirements.

5. There is an opportunity to produce about a 30% alignment between ROS and Research Fish by aligning the definition of similar fields and merging of vocabularies. There is a 50% discrepancy between the information requests made by councils to institutions that cannot be resolved by minor adjustments of existing fields. This is due in part to the different business needs of the councils and discipline-specific reporting requirements. Negotiations would be required between funders to carry out such alignment and was outside the scope of UKRISS.

6. The greatest current opportunities for harmonisation are around concrete output types such as publications. More subjective areas, where there is less consensus on what should be reported (eg impact), will require more work. However, reporting of impact is where the research councils and charities will gain the most benefits from harmonisation.

7. The project has developed software prototypes to demonstrate how fields in the core profile can be validated using external data sources and for compliance with the profile schema. The initial evaluation indicates that there is considerable scope for producing automated tools to improve the data quality of the research information at source.

8. Prototype aggregation tools have been developed that demonstrate the potential for combining and reusing research information for internal reporting and business analytics by both institutions and funders.

9. The UKRISS Crosswalk Connector demonstrates how manual effort in compiling reports for research funders can be reduced, and data quality improved, by enabling research information to be gathered from multiple institutional systems (eg HR, finance, CRIS) and mapped to the core profile. Funder-specific templates can be applied to filter the information to meet specific reporting requirements.

10. Adoption of the UKRISS core profile would be an important enabler for GtR and other consumers of research information (eg through G4HE).

4.2 Conclusions relevant to the wider community

1. Institutions would be major beneficiaries of the harmonisation of research information reporting. The benefits would primarily be accrued through more efficient reporting procedures, improved data quality and business intelligence analysis. Considerable effort for implementing harmonisation falls on funders. The short-term direct benefits for funders are unlikely to be as significant as for institutions. Indeed funders have already made investments in research outputs systems, which meet many of their most immediate needs. More of the work implied by not implementing harmonisation falls on institutions.

2. Governance of a harmonised profile as yet remains unresolved. The current proposal addresses primarily reporting to RCUK, and contributes to their internal harmonisation activities. However, there is a potential role for all organisations that collect research information to have a role in shaping a core profile, including institutions, statutory bodies and funders other than RCUK (including charity funders).

3. There was an identified overlap with the work of HEDIIP, particularly in the area of reporting on research students. HEDIIP also expressed an interest in using CERIF to model some aspects of teaching. Vocabularies are another area where synergies could be further exploited.

4. For many institutions RCUK funding represents only a fraction of the research funding (typically in the range of 30 to 50%). There are opportunities to extend the core profile to other funders, such as the European Commission, and charities, as well as to other forms of statutory reporting such as REF.

4.3 Conclusions relevant to Jisc

1. The progress on the harmonisation and exchange of research information reporting and implementation of CERIF in the UK HE sector has been due in large part to the support of Jisc. Currently there are no large scale production deployments of CERIF, and there are still challenges in implementing CERIF in production environments. Many stakeholders are still unconvinced that the benefits of using CERIF outweigh the considerable complexity. Continuation of the support for harmonisation and development of tools by Jisc is essential until CERIF can be proven in a large scale system, and wider adoption and benefits realisation is achieved.

2. There is a need to support the establishment of governance structures for harmonisation and to obtain buy-in from key stakeholders. Jisc is well placed to play a leading role in bringing key stakeholders together and articulating the benefits of such harmonisation for the sector, carrying forward the work done by UKRISS and previous Jisc-funded projects in a clear and coordinated way.

3. Use of CERIF supports the model of an open ecosystem (see Jisc charter), which gives vendors and suppliers the opportunity to build products which will interoperate and creates a marketplace for competition and innovation, from which the whole sector can benefit. Indeed, most commercial institutional CRIS systems already provide some degree of CERIF compliance.

4. The validation and aggregation tools developed in the project demonstrate the potential value of tools that could improve data quality and exploit open harmonised research information, leading to improvements in business intelligence and enhancing tasks such as business analysis at institutional and national levels. There is strong potential for further investigation and prototyping in this area.

5. There is clear evidence, through multiple presentations of UKRISS at international conferences, that the UK is an international leader in the investigation of CERIF and its application to increasing the exchange of research information.

5. Recommendations

This section describes the recommendations arising from the UKRISS work, using the same categories as in section 4.

5.1 Specific recommendations

1. Promote the standardisation of the core reporting profile and vocabularies through the CASRAI UK group or another forum with cross-sector representation [Research funders, institutions, charities and statutory bodies].

2. Investigate the closer business alignment of reporting fields in terms of their value versus the effort required to collect them [Research funders supported by institutions and statutory bodies].

Comment on recommendation 2: The alignment of business requirements for research reporting was outside the scope of UKRISS. However, we believe there is considerable scope for research councils to align and prioritise their reporting requirements, which would result in a further reduction in the burden on institutions.

3. Promote alignment between teaching and research information, particularly in the development of sector-wide vocabularies, and the area of research students, interacting with HEDIIP [RCUK, statutory bodies].

5.2 Recommendations for the wider community

1. Agree that a single organisation should be responsible for governance of the core profile [RCUK, institutions, statutory bodies].

Comment on recommendation 1: Development of the core profile can only be achieved in small steps. Furthermore, reporting requirements change over time, so the profile would need to evolve and be managed on an ongoing basis. Institutions as well as funders should be involved in this activity, as there is a requirement to align the reporting needs of funders with the capabilities of institutions to collect such information.

2. Consider adoption of the core profile (or portions of it) by non-RCUK funders and statutory bodies. In particular, the

UKRISS core profile should be considered as an important element of future REFs [Other funders including charities, statutory bodies].

5.3 Recommendations for Jisc

1. Arrange a meeting of senior representatives of the key stakeholders, including RCUK funders and institutions to review and follow up the UKRISS findings and to carry forward the proposals for harmonisation.

2. Promote the development of the core profile through support for the CASRAI UK task groups and other initiatives. This includes continuing support for CASRAI over a timeframe that enables consensus to be reached.

3. Promote the development of an open ecosystem of research information systems based on CERIF by facilitating engagement between stakeholders and funding of further pilots and studies.

4. Promote the prototyping and adoption of registries for identifiers to improve data quality and interoperability.

5. Conduct a more comprehensive cost-benefit analysis to build on the benefits, risk and measurements identified in the UKRISS preparation for a full business case for harmonisation. A prerequisite would be buy-in from both funders and institutions.

6. Investigate the applicability of UKRISS techniques to address analogous issues in the return of teaching and other non-research information from institutions to external bodies.

6. Implications for the future

In this section we consider the implications of the UKRISS work for the key stakeholders, and how the work can be built on and extended in the future.

6.1 Implications for stakeholders

6.1.1 Institutions

For institutions, UKRISS has provided an understanding of the potential benefits of harmonisation of research information reporting:

- Simplification of reporting requirements and systems by increasing the number of reporting fields that are common to multiple funders
- Increased access to reporting data that can be compared across multiple funders
- Automation of reporting processes in extracting and compiling reporting information and mechanism to improve data quality
- Potential to develop business intelligence and aggregation tools
- Greater compliance with requirements to submit timely returns

Institutions are dependent on funders on one hand to implement the UKRISS recommendations, and on the continued development of the tools and services demonstrated within the UKRISS project.

6.1.2 Researchers

There are three main potential benefits of implementation of the UKRISS proposals for researchers:

- Reduce duplication of effort in entering information into multiple systems
- Potential to provide higher quality resources to support research through access to high quality research information
- Greater chance that institutions can do the reporting on behalf of the researchers, and hence reduce the burden on them

6.1.3 RCUK funders

For RCUK funders, UKRISS has provided aggregated and core reporting profiles that would enable closer harmonisation of data entered into the ROS and Research Fish reporting systems. In order to demonstrate the value of this, the project has developed software prototypes and an outline of the business case. In particular, this analysis lists the benefits and risks involved. A more comprehensive analysis of the costs involved would require more detailed input from the councils and other funders.

In the short term, the UKRISS work provides input to the review of the existing RCUK reporting systems at a business and technical level. In the longer term, it highlights the potential benefits, not only for the councils themselves but across the wider sector.

6.1.4 Other research funders including charities

Although charity funders were not in scope of the work on the core profile definition, there is considerable overlap in the information that is reported between RCUK and non-RCUK funders. The UKRISS work enables charities to assess where there is scope to align their reporting with a harmonised profile and to understand the benefits of doing so. For the business case for charities to be articulated, more detailed input from the charities would be required.

6.1.5 Implications for government and statutory bodies

The government has a strong interest in increasing efficiency across the sector, both in terms of reducing costs as well as promoting the competitiveness of the UK research base. In particular through the BIS GtR project, there is the intention to make the results of publicly funded research available more widely to support greater efficiency within the HE sector as well as to promote the reuse of academic research by industry to support economic growth. The UKRISS work contributes directly a means of greater harmonisation of the information gathered as well as to improvements in data quality.

Statutory bodies such as HEFCE and HESA are responsible for data collection which overlaps with the data already collected by the RCUK systems. There is a potential therefore to align the HE-BCI survey and also future REF exercises with the UKRISS core profile definitions.

6.1.6 Implications for vendors

Vendors were seen as key stakeholders in the longer term, as they are in a position to incorporate UKRISS features into their products, securing commercial advantage and user benefits (eg business intelligence).

6.1.7 Implications for CERIF community

The UKRISS project has contributed to the CERIF community in a number of ways:

- Provided tools for CERIF
- Provided a reference approach and model to other countries facing similar issues
- Developed proof of concept software and valuable CERIF XML examples following real-world requirements
- Produced proposals for the application and standardisation of requirements-driven business rules, ie recommendation as to how to define and specify, in general, object boundaries beyond just representations and descriptive vocabularies

6.2 Further technical development

There are many ways in which the results of UKRISS could be extended through further technical development work. These include:

1. Further development of validation tools – even without harmonisation, these tools can have significant impact in the community, and be able to make data easier to interchange even in the absence of the UKRISS models.
2. Development of aggregation user stories and analytical dashboards (along the lines of G4HE) to answer questions valuable to the community (and in particular funders and HEIs).
3. Alignment of the UKRISS core profile with the mappings and vocabularies being developed by GtR.
4. Further development and evaluation of the Crosswalk Connector within institutional and funder environments. This might be integrated with tools to enable funders to harvest research information from institutions automatically, and also to automate the dissemination of information from funders to institutions.
5. Development of open publications repositories for a wider set of subject areas outside of medical research, based on the model of PubMed.
6. The UKRISS core profile is available to be extended and developed further for other applications. The UKRISS project already provided a contribution to the Jisc-funded DESCRIBE project on research impact (an area which is still under development). The EPSRC-funded Equipment.Data project (http://equipment.data.ac.uk) has expressed an interest in extending the profile to include reporting on use of research equipment and facilities by research projects.

6.3 Sustainability

The following are the main ways in which we plan that the sustainability of the results from UKRISS will be achieved:

- The UKRISS core profile, supporting analysis will be presented to the CASRAI UK task group for further discussion with interested stakeholders. We anticipate that the RCUK will play a leading role in these technical discussions. This process should be supported by Jisc
- The UKRISS core profile will also be of value for the GtR project, as the UKRISS work covers many output types that have not yet been implemented in GtR
- RCUK is currently reviewing its business requirements for reporting and the supporting research outcomes systems.

The results of UKRISS, in particular the core profile, business case and data validation tools will provide an important input to this activity

- The UKRISS Steering Board contains representatives of the key stakeholders. The Steering Board has been involved in providing specialist inputs, shaping the project goals and strategy, as well as being kept regularly up-to-date with progress. Hence it is well placed to act as advocate for the UKRISS work within their respective organisations
- HEDIIP has a role in sustaining the findings of UKRISS through alignment of reporting requirements and vocabularies between the teaching and research areas. There has already been collaboration between HEDIIP and UKRISS through work on a report and several face-to-face meetings
- From an international perspective, as a UKRISS project partner, euroCRIS will be able to provide continuing support for disseminating the project results to a wider audience and promoting harmonisation activities around CERIF at an international level, based on the experiences of UKRISS
- The partner institutions Brunel, Exeter and King's College London are planning to continue the collaboration on the Crosswalk Connector beyond the end of UKRISS
- Vendors have a strong interest in a harmonised reporting profile and have an important role in moving the core profile proposals into a production environment. The software prototypes developed in UKRISS such as the data validation and aggregation tools, which are available as open source, provide guidance for implementation, both for using the core profile as well as the potential applications that can be built upon it

6.4 Long term project contact

The primary long term contact for enquiries relating to the UKRISS project is:

Dr Simon Waddington
Centre for e-Research
King's College London
26-29 Drury Lane
London WC2B 5RL.
Email: **simon.waddington@kcl.ac.uk**

7. References

Bolton S. (2010). Business case for the adoption of a UK standard for research information interchange. Report to Jisc. 2010. Web. 9 September 2013. www.jisc.ac.uk/publications/reports/2010/businesscasefinalreport

BRUCE Project (2012). Web. 3 July 2014. http://bruceatbrunel.wordpress.com

CASRAI (2014). Web. http://casrai.org

Common European Research Information Format (CERIF) (2014). Web. 26 June 2014. www.eurocris.org/Index.php?page=CERIFreleases&t=1#

CrossRef (2013). Web. 9 September 2013. www.crossref.org

DESCRIBE – Definitions, Evidence, and Structures to Capture Research Impact and Benefits – Project. University of Exeter (2013). Web. 26 June 2014. www.exeter.ac.uk/media/universityofexeter/research/inspiringresearch/describeproject/pdfs/2013_06_04_Executive_Summary_and_Recommendations_FINAL

Duryea M., Hochman M., Parfitt A. (2007). Measuring the impact of research. February 2007, Research Global. Web. 6 January 2014. www.atn.edu.au/Documents/Articles/2011/2010/2009/2008/2007/Measuring%20the%20impact%20of%20research

Elsevier Pure (2013). Web. 9 September 2013. http://info.scival.com/pure

ePrints repository system (2013). Web. 6 September 2013. www.eprints.org

FundRef funder identification service (2013). Web. 6 September 2013. www.crossref.org/fundref/index.html

Gateway to Research (2013). Web. 9 September 2013. www.rcuk.ac.uk/research/Pages/gtr

HEDIIP report, The HE Information Landscape: Creating and Managing a Data Model (2013). Web. 6th January 2014. www.hediip.ac.uk/wp-content/uploads/HEDIIP_Data_Language_Report_2013-09

HESA Information Landscape Study (2013). Web. 6th January 2014. http://landscape.hesa.ac.uk

Measuring Impact under CERIF (MICE) project final report (2011). Web. 26 June 2014. http://mice.cerch.kcl.ac.uk/wp-uploads/2011/06/ImpactUnderCERIF

ORCID (2013). Web. 6 September 2013. https://orcid.org

REF Impact Pilot Exercise: Findings of the expert panels. A report to the UK higher education funding bodies by the chairs of the impact pilot panels (2010). Web. 6 January 2014. www.ref.ac.uk/media/ref/content/pub/researchexcellenceframeworkimpactpilotexercisefindingsoftheexpertpanels/re01_10

Ringgold (2014). Web. 4 January 2014. www.ringgold.com

RiO Extension Project (2012). Web. 5 September 2013. www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/repositories/rioextension

Rogers N., Huxley L., Ferguson N. (2009). Exchanging Research Information in the UK. EXRI-UK: a study funded by Jisc. 2009. Web. 6 September 2013. http://repository.jisc.ac.uk/448

Symplectic Elements (2013). Web. 6 September 2013. www.symplectic.co.uk

Thomson Reuters Converis (2013). Web. 10 March 2014. www.converis5.com

UKRISS CERIF Elements and Vocabularies Landscape Survey (2013). Web. 26 June 2014. http://ukriss.cerch.kcl.ac.uk/cerif-elements-and-vocabularies-landscape-survey

UKRISS Feasibility Study (2013). http://goo.gl/4fzCXH

UKRISS Landscape Study (2013). Web. 20 January 2013. http://ukriss.cerch.kcl.ac.uk/?p=75

Waddington S., Sudlow A., Walshe K., Scoble R., Mitchell L., Jones R., Trowell S. (2013). Feasibility Study Into the Reporting of Research Information at a National Level Within the UK Higher Education Sector. New Review of Information Networking Volume 18, Issue 2, 2013, pp74-105. DOI:10.1080/13614576.2013.841446. www.tandfonline.com/eprint/76QMhvSWJTEZMcghVewh/full#.Unt_wV9FCUk

8. Glossary

Table 8.1 provides a glossary of the key terminology used in this document.

Term	Description
HEI	Higher Education Institution
CERIF	Common European Research Information Format
HE-BCI Survey	Higher Education Business and Community Interaction Survey
RCUK	Research Councils UK
ROS	Research Outcomes System (RCUK)
RF	Research Fish
RO	Research Organisation
HEFCE	Higher Education Funding Council for England
HESA	Higher Education Statistics Agency
HEDIIP	Higher Education Data & Information Improvement Programme
CASRAI	Consortia Advancing Standards in Research Administration Information

Table 8.1: Glossary of terms

9. Appendix 1: UKRISS Steering Board

The membership of the UKRISS Steering Board is shown in table 9.1.

Name	Affiliation
Ian Carter (Chair)	University of Sussex, ARMA
Maja Maricevic	British Library
Liz Philpots	AMRC
Geraldine Clement-Stoneham	MRC
Gerry Lawson	EPSRC/GtR
Luke Moody	ESRC
Geoff Rodgers	Brunel University
Neil Jacobs	Jisc
Kimberley Hackett	HEFCE
Kevin Dolby	Wellcome
Ian McArdle	Imperial College, University of London
Luke Taylor	University of Bristol, UCISA
Verena Weigert	Jisc
Gregor McDonagh	NERC

Table 9.1: Steering Board membership

10. Appendix 2: Core Information Reporting Profile

The UKRISS Core Information Reporting Profile is composed of the upper level reporting types such as "Research Output", "Research Transfer", "Research Outcome", "Research Impact" and "Measurement". These have been understood as follows:

- **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 outputs; inventions 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 as presented in figure 10.1 is reflected (modelled) in an Excel sheet (Appendix 4). The reporting types are subsumed by record types; eg Publication is subsumed under Research Output, Spin-Out and Further Funding under Research Outcome.

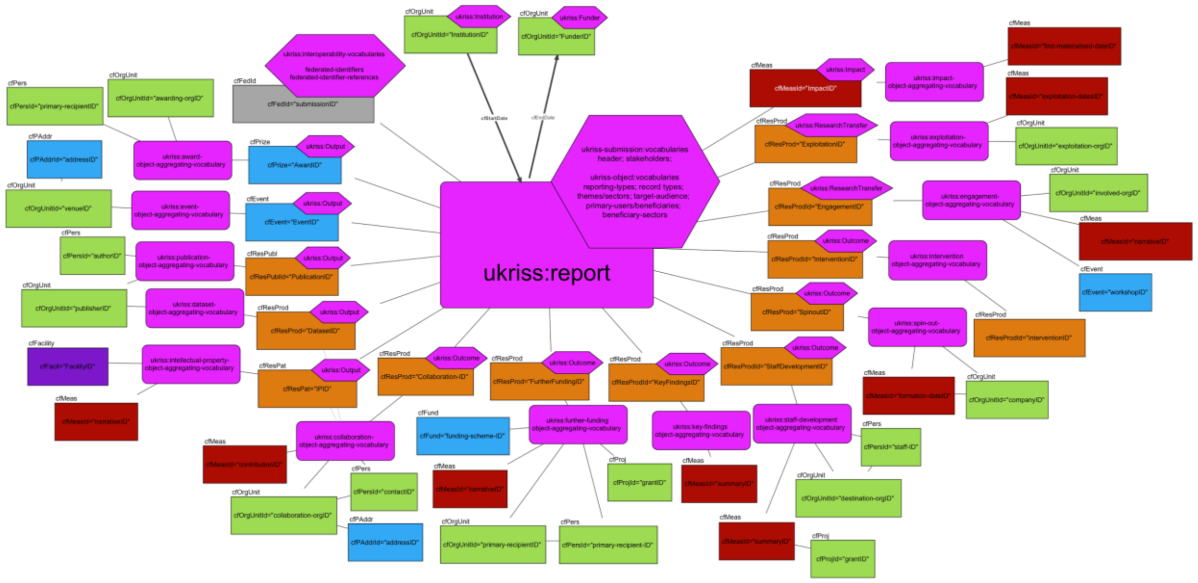

cfOrgUnit
ukriss:Institution
cfOrgUnitId="InstitutionID"
cfOrgUnit
ukriss:Funder
cfOrgUnitId="FunderID"
cfStartDate
cfEndDate
ukriss:interoperability-vocabularies
federated-identifiers
federated-identifier-references
cfFedId
cfFedId="submissionID"
cfPers
cfPersId="primary-recipientID"
cfOrgUnit
cfOrgUnitId="awarding-orgID"
cfPAddr
cfPAddrId="addressID"
ukriss:award-object-aggregating-vocabulary
cfPrize
ukriss:Output
cfPrize="AwardID"
cfOrgUnit
cfOrgUnitId="venueID"
ukriss:event-object-aggregating-vocabulary
cfEvent
ukriss:Output
cfEvent="EventID"
cfPers
cfPersId="authorID"
ukriss:publication-object-aggregating-vocabulary
cfResPubl
ukriss:Output
cfResPublId="PublicationID"
cfOrgUnit
cfOrgUnitId="publisherID"
ukriss:dataset-object-aggregating-vocabulary
cfResProd
ukriss:Output
cfResProd="DatasetID"
cfFacility
cfFacil="FacilityID"
ukriss:intellectual-property-object-aggregating-vocabulary
cfResPat
ukriss:Output
cfResPat="IPID"
cfMeas
cfMeasId="narrativeID"
ukriss-submission:vocabularies
header; stakeholders;
ukriss-object:vocabularies
reporting-types; record types;
themes/sectors; target-audience;
primary-users/beneficiaries;
beneficiary-sectors
ukriss:report
cfMeas
ukriss:Impact
cfMeasId="ImpactID"
ukriss:impact-object-aggregating-vocabulary
cfMeas
cfMeasId="first-materialised-dateID"
cfMeas
cfMeasId="exploitation-datesID"
cfResProd
ukriss:ResearchTransfer
cfResProd="ExploitationID"
ukriss:exploitation-object-aggregating-vocabulary
cfOrgUnit
cfOrgUnitId="exploitation-orgID"
cfResProd
ukriss:ResearchTransfer
cfResProdId="EngagementID"
ukriss:engagement-object-aggregating-vocabulary
cfOrgUnit
cfOrgUnitId="involved-orgID"
cfMeas
cfMeasId="narrativeID"
cfEvent
cfEvent="workshopID"
cfResProd
ukriss:Outcome
cfResProdId="InterventionID"
ukriss:intervention object-aggregating-vocabulary
cfResProd
ukriss:Outcome
cfResProd="SpinoutID"
ukriss:spin-out-object-aggregating-vocabulary
cfResProd
cfResProdId="interventionID"
cfMeas
cfMeasId="formation-dateID"
cfOrgUnit
cfOrgUnitId="companyID"
cfResProd
ukriss:Outcome
cfResProd="Collaboration-ID"
cfResProd
ukriss:Outcome
cfResProd="FurtherFundingID"
cfResProd
ukriss:Outcome
cfResProdId="KeyFindingsID"
cfResProd
ukriss:Outcome
cfResProdId="StaffDevelopmentID"
ukriss:collaboration-object-aggregating-vocabulary
cfMeas
cfMeasId="contributionID"
cfPers
cfPersId="contactID"
cfOrgUnit
cfOrgUnitId="collaboration-orgID"
cfPAddr
cfPAddrId="addressID"
cfFund
cfFund="funding-scheme-ID"
ukriss:further-funding object-aggregating-vocabulary
cfMeas
cfMeasId="narrativeID"
cfOrgUnit
cfOrgUnitId="primary-recipientID"
cfPers
cfPersId="primary-recipient-ID"
cfProj
cfProjId="grantID"
ukriss:key-findings object-aggregating-vocabulary
cfMeas
cfMeasId="summaryID"
ukriss:staff-development object-aggregating-vocabulary
cfPers
cfPersId="staff-ID"
cfOrgUnit
cfOrgUnitId="destination-orgID"
cfMeas
cfMeasId="summaryID"
cfProj
cfProjId="grantID"

Figure 10.1: UKRISS Core Research Information Reporting Profile in CERIF including aggregations representing the use case "Institution submits final report to funder"

CERIF is an open data model supplying a formal syntax and declared semantics, thus enabling the implementation of any structure and aggregation of defined objects; it allows for time-stamped multiple-type or multiple-role assignments. The UKRISS bottom-up extraction of fields from within the underlying investigated systems provided clear reporting object boundaries.

The identified UKRISS reporting objects have been presented to the euroCRIS community and to the CERIF TG and members. Due to their global relevance they have been considered a very valuable output for much wider and further reuse beyond UKRISS and even the UK. That is, the UKRISS reporting objects contribute both conceptually (object-boundaries and implied vocabulary usage) as well as formally or technically towards CERIF modelling and mappings.

A very rough comparison analysis at object level revealed the overlap in reporting objects within UKRISS and the German Research Core Dataset project, where a discussion and further comparison are being considered even beyond UKRISS, to estimate the overlap and potential degree of object-reuse within this use case while being aware of the different modelling and analysis approaches.

The bottom-up analysis revealed a semantic gap and a number of duplicate or missing fields across objects. The introduction of an upper reporting level and the use case guided the UKRISS model developments towards semantic balance as well as with respect to the normalisation of reporting objects and enhancements through generic elements for increased consistency. That is, some elements are required at least once with each reporting object record in addition to object describing attributes:

- ukriss-submission:header (eg ID; version; Date; Contact; Type)
- ukriss-submission:stakeholders (eg ukriss:funder)
- ukriss-object:reporting-types (eg ukriss:research-output)
- ukriss-object:record-types (eg ukriss:publication)
- ukriss-object:themes-sectors (eg ros:aerospace-defence; rf:health-care)
- ukriss-object:target-audience (eg ros/rf:public)
- ukriss-object:primary-users-beneficiaries (eg ros:academic-institution)
- ukriss-object:beneficiary-sectors (eg ros:innovation; rf:private)
- ukriss:url (eg http://gateway-to-research.ac.uk/ID)
- ukriss:federated-identifiers (eg cerif:doi; cerif:ukpmc; cerif:issn; cerif:isbn)
- ukriss:federated-identifier-references (eg ukriss:orcid; ukriss:ukprn)

The introduced object and record extensions ensure compliance with the use case and contextual reuse and re-identification of objects. The proposed elements and corresponding vocabulary terms are available in the UKRISS Core Information Reporting Profile Model, and are reflected in CERIF XML examples and reporting object visualisations as provided within Appendix 4.

The introduction of the use case "Institution submits final report to funder" as indicated in section 3.3.1 guided the modelling of the UKRISS reporting objects in CERIF, ie the CERIF mappings. The Excel sheet provides the mappings for each underlying UKRISS object; its describing fields and vocabularies. Where the Research Output objects are all explicitly available as entities in the CERIF model, this is not the case with Research Transfer objects or Research Outcome objects. These have thus been modelled as research products in the first instance, and secondly linked to their more functional entities. For example, Staff Development is at first considered an outcome, ie cerif:product and second a cerif:person and its aggregates. The same holds for Collaboration, first considered an outcome, ie cerif:product and second a cerif:organisation and its aggregates (see also images in the remainder and CERIF XML example files).

The entire UKRISS Core Information Reporting Profile is available in Appendix 4, including the potential and recommended harmonisation changes (counts). That is, the sheet includes the upper reporting level and generic field extensions, the proposed identifier changes and thus normalisation suggestions as well as vocabulary harmonisation suggestions and object boundaries. It supports the CERIF structure aimed at reduction of field duplication (error-prone string fields) towards more identifier-driven and semantically (or conceptually)-aligned interoperability and thus towards improved information quality and reuse.

The modelling of the UKRISS reporting objects did not require 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/triggering following eg requirements or "business" rules. This proposal has also been forwarded to the CERIF TG for further consideration during the Porto TG meeting in November 2013 and discussions will continue with interested parties.

For further details on the UKRISS Core Profile, see Appendix 4.

11. Appendix 3: Harmonisation analysis

The UKRISS model is aimed at providing guidance for harmonisation of ongoing reporting activities at three levels: with objects, at object fields, and with field inherent vocabularies.

11.1 Harmonisation analysis at object level

Figure 3.1 presents the UKRISS modelling approach and uses colour coding to illustrate the degree of object harmonisation. Different shapes are used to distinguish the reporting objects. Current harmonisation is indicated with green objects while potential harmonisation objects are shaded pink, anticipating implementations through the developed UKRISS Core Information Reporting profile. The lines drawn from the bottom indicate the underlying systems of the reporting objects, where eg Exploitation and Key Findings are derived from ROS, while eg Product/Intervention is derived from Research Fish, and Event and Dataset are proposed objects not inherent in the underlying system as individual objects but considered relevant within the developed UKRISS model. A potential object harmonisation at the upper reporting level is shown by pink lines to the introduced reporting types: Research Output, Research Transfer, Research Outcome, Research Impact and Measurement.

11.2 Harmonisation at field/vocabulary level

The bottom-up analysis uncovered a semantic gap between the objects' field labels and definitions or descriptions. These included duplications of fields across objects as well as a lack of semantic clarity of labels, descriptions or definitions. Furthermore, objects differed in the number of describing fields, ie descriptive granularity. The reasons for these differences are probably to be found in the ROS and RF implementation histories. But, possibly they also result from the different disciplines they serve and support, and are certainly also related to the "age" of or the knowledge about the reporting objects themselves. For example, "publication" is much better understood and thus more extensively described or represented (field counts) than any other investigated reporting object (see figure 3.2).

The ROS and RF histories have not and will not be further investigated within the UKRISS project. The estimated degree of current/potential harmonisation between the two systems has been semantically validated with sample data records.

Figure 3.2 presents the inherent ROS and Research Fish reporting objects with the container object "Other Research Output" on the left, under which Dataset and Event are non-explicitly subsumed. It illustrates the degree of harmonisation for the inherent individual reporting objects in ROS and RF, revealing again that, for example, Key Findings and Exploitation are objects inherently reported through ROS, while Product/Intervention is reported through RF. Likewise, the field counts at object level demonstrate the difference in descriptive granularity at object level. For example, Publication counts highest in ROS and RF, while Intellectual Property or Spin-Out objects count a few fields only. Overall it is obvious that the reporting objects in ROS are described through more fields than in RF.

The first investigation of "semantic similarity" at field and vocabulary label levels did not consider the REF and HE-BCI objects and elements. These have been taken into account with the more conceptual approach of profile modelling and subsequent CERIF mappings. That is, the possibility of (sub-)object and element reuse within the proposed reporting profile including vocabularies and reporting elements following the use case. Additional vocabularies have been collected through the vocabulary landscape study and beyond, as presented in the next sub-section. The final degree of potential harmonisation at field and vocabulary level through the proposed UKRISS model anticipating a normalised CERIF-driven implementation is not calculated entirely, but only demonstrated by example counts (see figure 3.4) from counts within the Excel sheet (Appendix 4).

11.3 Vocabulary landscape study

In addition to the vocabularies extracted from ROS and RF, vocabularies from relevant UKRISS-external initiatives were collected in the UKRISS CERIF Elements and Vocabularies Landscape Survey (2013), namely RIOXX, GtR, CASRAI, Pure, CERIF for Datasets (C4D) and IRIOS. The relevant elements of these vocabularies were added to the final collection for further investigation beyond UKRISS. Such vocabularies could contribute to harmonisation at the "global" scale. Furthermore, the ISO codes for languages, countries and currencies, and the REF and HE-BCI vocabulary elements have been added to the collection.

The final UKRISS vocabulary collection comprises a total of more than 100 vocabularies – that is, role and type schemes subsuming terms. The collection enables a UK reporting landscape harmonisation through the UKRISS Core Information Reporting Profile following the CERIF XML data model structure. The vocabulary collection is to a large extent comprised of vocabularies and terms extracted from ROS and RF (merged within UKRISS) – the vocabulary source is reflected in the prefixes.

11.3.1 Extracted ROS/RF vocabularies

- ukriss-submission:header
- ukriss-submission:stakeholders
- ukriss-object:federated-identifiers
- ukriss-object:record-types
- ukriss-object:reporting-types
- ukriss:-object:interoperability
- ukriss-object:primary-users/beneficiaries
- ukriss-object:target-audience
- ukriss-object:themes/sectors
- ukriss-object:beneficiary-sectors
- ukriss:organisation-types
- ukriss:countries
- ukriss:narratives/comments/figures
- ukriss-output:other-types
- ukriss-output:states
- ukriss-output:target-audience
- ukriss-output:person-output-roles
- ukriss-output:organisation-output-roles
- ukriss-output:dates
- ukriss-output:embargo-periods
- ukriss-publication:types
- ukriss-publication:states

- ukriss-publication:-(open)-access-types
- ukriss-dataset:types
- ukriss-event:types
- ukriss-event:duration-types
- ukriss-event:income
- ukriss-event:venue-types
- ukriss-event:geographic-areas
- ukriss-event:audience-size-types
- ukriss-collaboration:types
- ukriss-collaboration:roles
- ukriss-collaboration:states
- ukriss-further-funding:types
- ukriss-further-funding:funder-types
- ukriss-key-findings:types
- ukriss-staff-development:qualification-gained
- ukriss-staff-development:types
- ukriss-staff-development:project-roles
- ukriss-staff-development:destination-roles
- ukriss-staff-development:employment-roles
- ukriss-staff-development:secondment-roles
- ukriss-staff-development:fellowship-roles
- ukriss-staff-development:consultancy-roles
- ukriss-staff-development:employment-types
- ukriss-staff-development:destination-themes/sectors
- ukriss-staff-development:destination-countries
- ukriss-staff-development:number-ftes
- ukriss-exploitation:types
- ukriss-spin-out:types
- ukriss-ip:types
- ukriss-ip:disclosure-states
- ukriss-ip:location-filed
- ukriss-intervention:types
- ukriss-intervention:active-development-stages
- ukriss-intervention:development-states
- ukriss-dataset:types
- ukriss-engagement:types
- ukriss-engagement:states
- ukriss-object:target-audience
- ukriss-engagement:broadcast media
- ukriss-engagement:international-audience
- ukriss-engagement:public-scheme
- ukriss-award-recognition:types
- ukriss-impact:types

- ukriss-impact:influence-types
- ukriss-impact:influence-areas
- ukriss-impact:geographic-extent
- ukriss-impact:states

11.3.2 Relevant external vocabularies

- ref-common-mandatory-fields
- ref-organisation-types
- ref-multiple-submissions
- ref-action-types
- ref-contact-types
- ref-identifier-types
- ref-person-names
- ref-contract-types
- ref-sensitivity-types
- ref-output-types
- ref-units-of-assessment
- hebcis-research-related-activities/collaborative-research-involving-public-funding
- hebcis-research-related-activities/contract-research
- hebcis-business-and-community-services/consultancy-contracts
- hebcis-business-and-community-services/facilities-and-equipment-related-services-organisations-involved-and-income
- hebcis-business-and-community-services/courses-for-business-and-the-community
- hebcis-regeneration-and-development-programmes/programme
- hebcis-intellectual-property/disclosures-and-patents-filed-on-behalf-of-the-hei
- hebcis-intellectual-property/licence-numbers
- hebcis-intellectual-property/ip-income
- hebcis-intellectual-property/spin-of-activity
- hebcis-intellectual-property/spin-of-activity/numbers
- hebcis-social-community-and-cultural-engagement/designated-public-events
- hebcis-social-community-and-cultural-engagement/numbers-free-events
- hebcis-social-community-and-cultural-engagement/numbers-chargeable-events
- rioxx-application-profile-v1.0
- gateway-to-research-types
- casrai-research-personnel-profile/contributions/outputs
- research-personnel-profile/contributions/other-outputs
- pure-research-outputs
- cerif-4-datasets
- irios
- iso-country-codes
- iso-language-codes
- iso-currency-codes

The entire UKRISS vocabulary collection is available in the Excel sheet in Appendix 4 as part of the UKRISS Core Information Reporting Profile model. The spreadsheet contains the extracted and external as well as proposed potential harmonisation terms behind the above presented schemes, such as for example ros:schools/students; rf:schools. In addition, it contains the counts behind the proposed merging, normalisation or relocation of initial vocabulary occurrences within the UKRISS Profile.

Aggregation vocabularies to define the boundaries of the identified UKRISS reporting objects (sub-profiles) were part of the vocabulary landscape study and need a different consideration beyond the UKRISS project. Initial thoughts about rules and constraints have been forwarded to the CERIF TG for further discussion. The below images indicate the range and application of terms within the drawn boundaries of each UKRISS reporting object.

11.3.3 UKRISS reporting object aggregation vocabularies

CERIF allows for a representation of objects through aggregation of entities, their relationships and vocabularies in a timely manner. However, it does not supply rules to define the explicit extent, ie the boundaries or application profiles of a particular object. The images that follow visualise the aggregation of applicable CERIF elements with UKRISS reporting objects as defined in the UKRISS Core Information Reporting Profile (Appendix 4), where a formal boundary definition is considered beyond the scope of the UKRISS project, but where the experience from UKRISS will be a valuable input for continued related discussions in the CERIF TG.

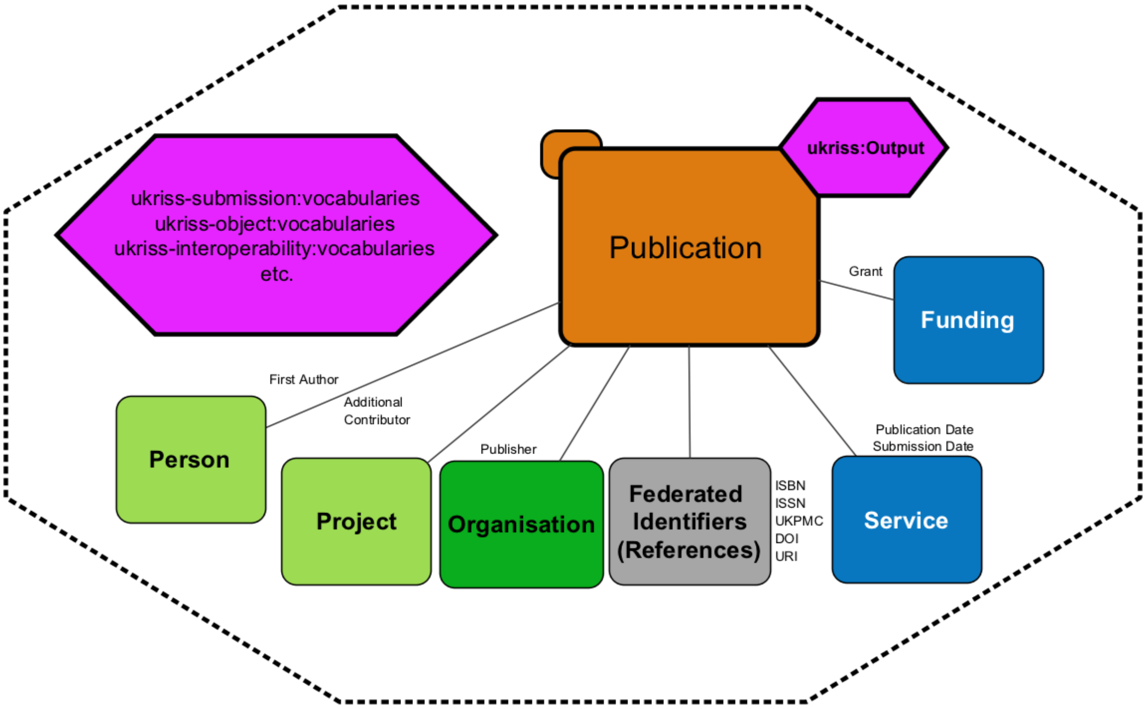


Figure 11.1: Aggregation of Publication elements through CERIF entities

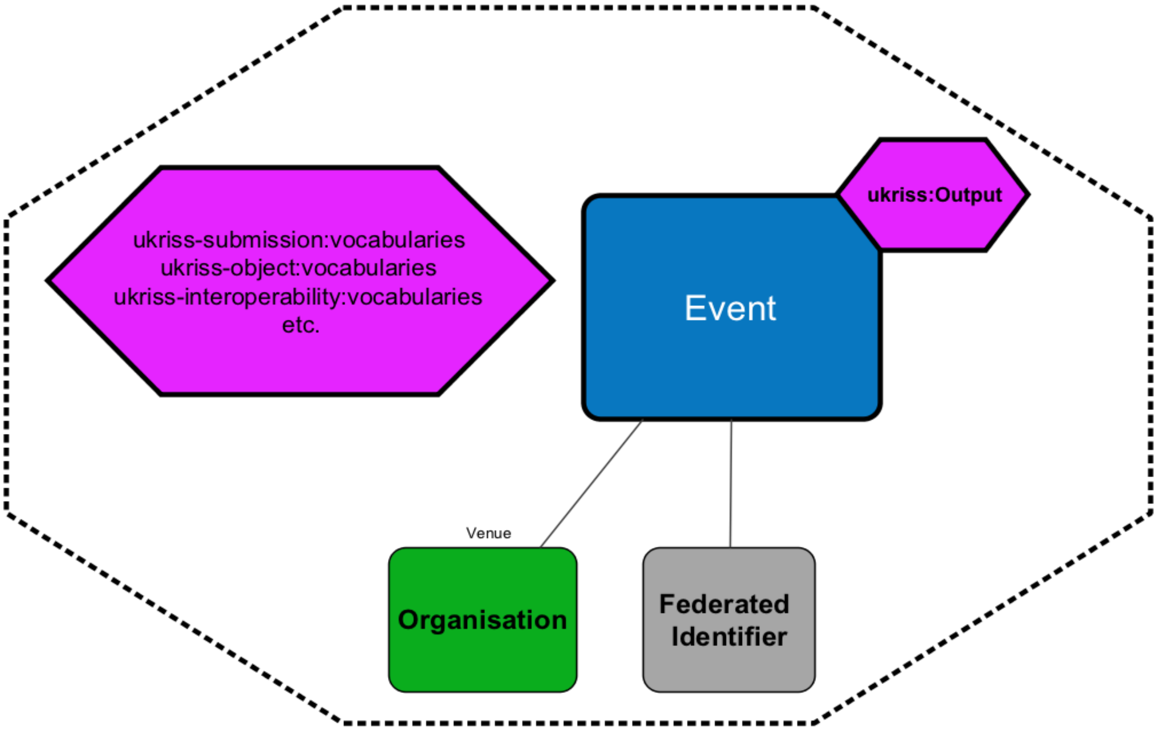

ukriss-submission:vocabularies
ukriss-object:vocabularies
ukriss-interoperability:vocabularies
etc.
ukriss:Output
Event
Venue
Organisation
Federated Identifier

Figure 11.2: Aggregation of Event elements through CERIF entities

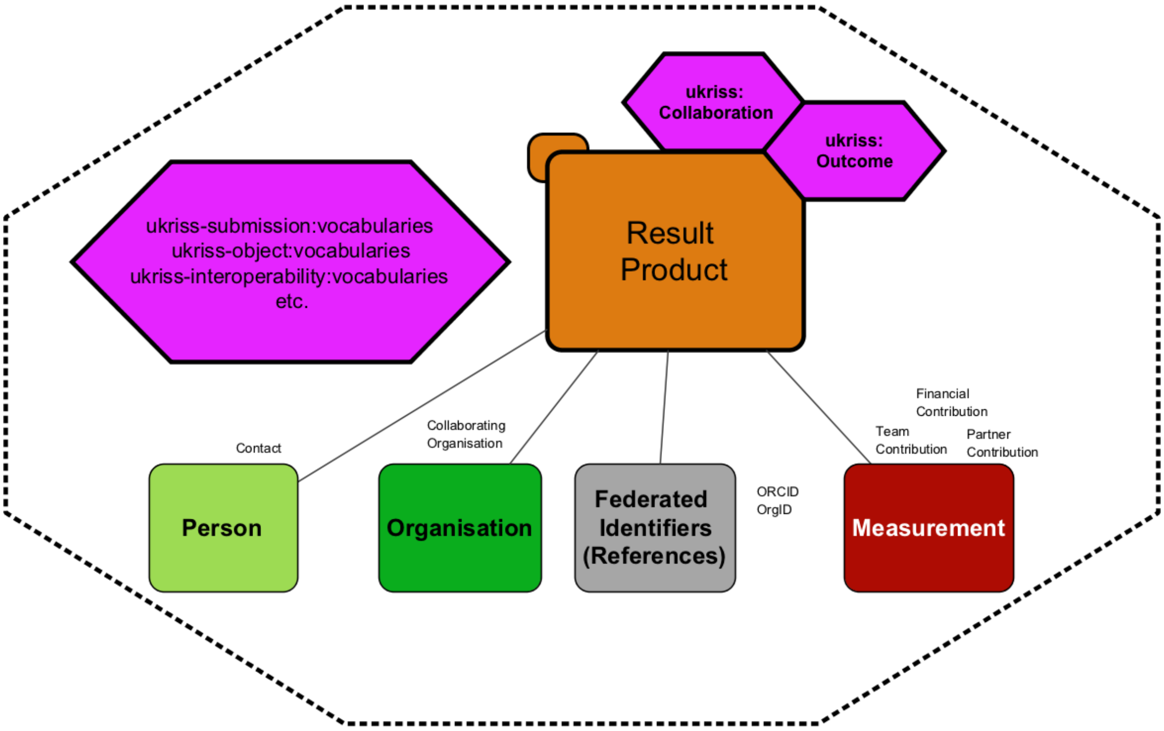


Figure 11.3: Aggregation of Collaboration elements through CERIF entities

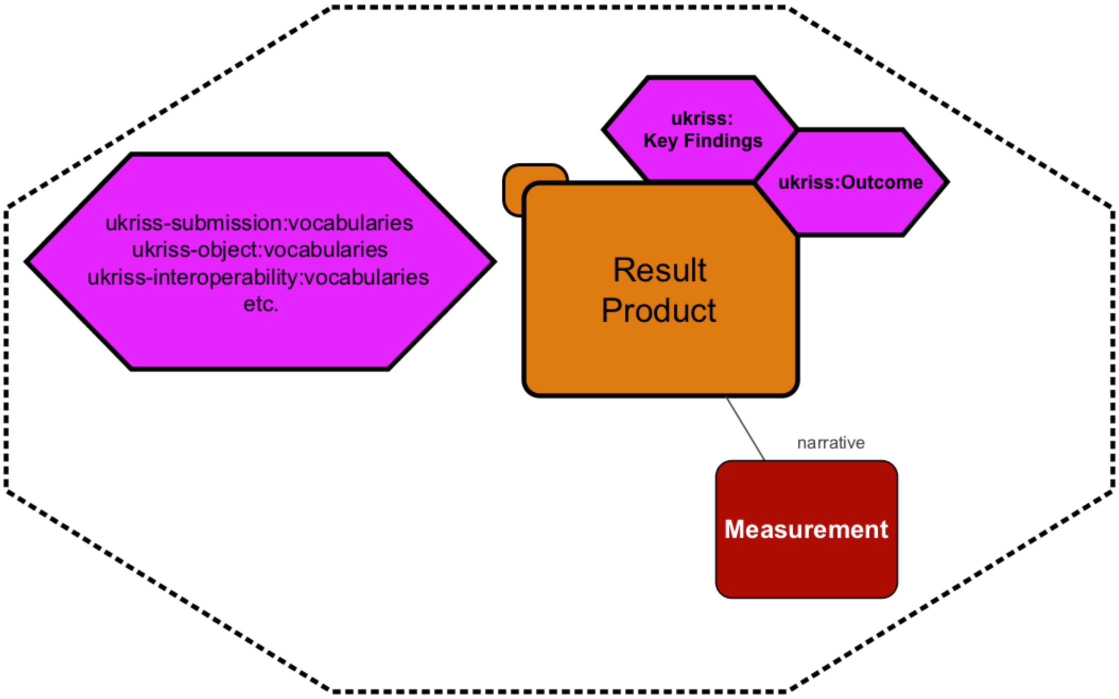

ukriss-submission:vocabularies
ukriss-object:vocabularies
ukriss-interoperability:vocabularies
etc.
ukriss:
Key Findings
ukriss:Outcome
Result
Product
narrative
Measurement

Figure 11.4: Aggregation of Key Findings elements through CERIF entities

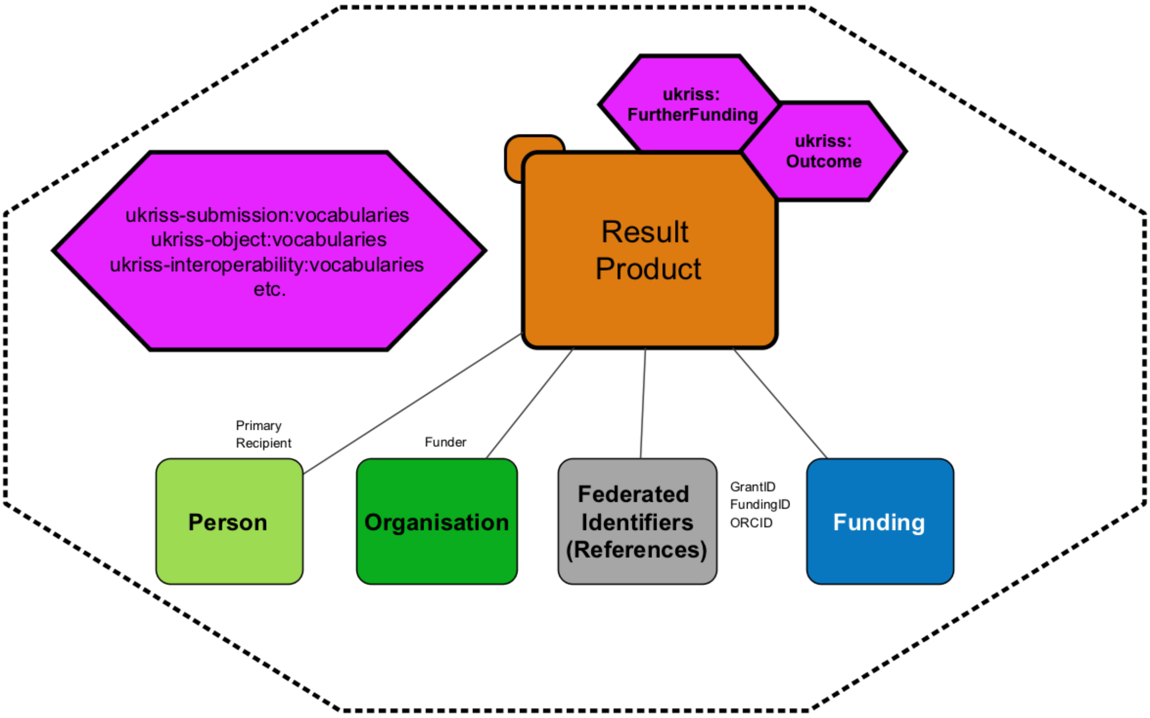


Figure 11.5: Aggregation of Further Funding elements through CERIF entities

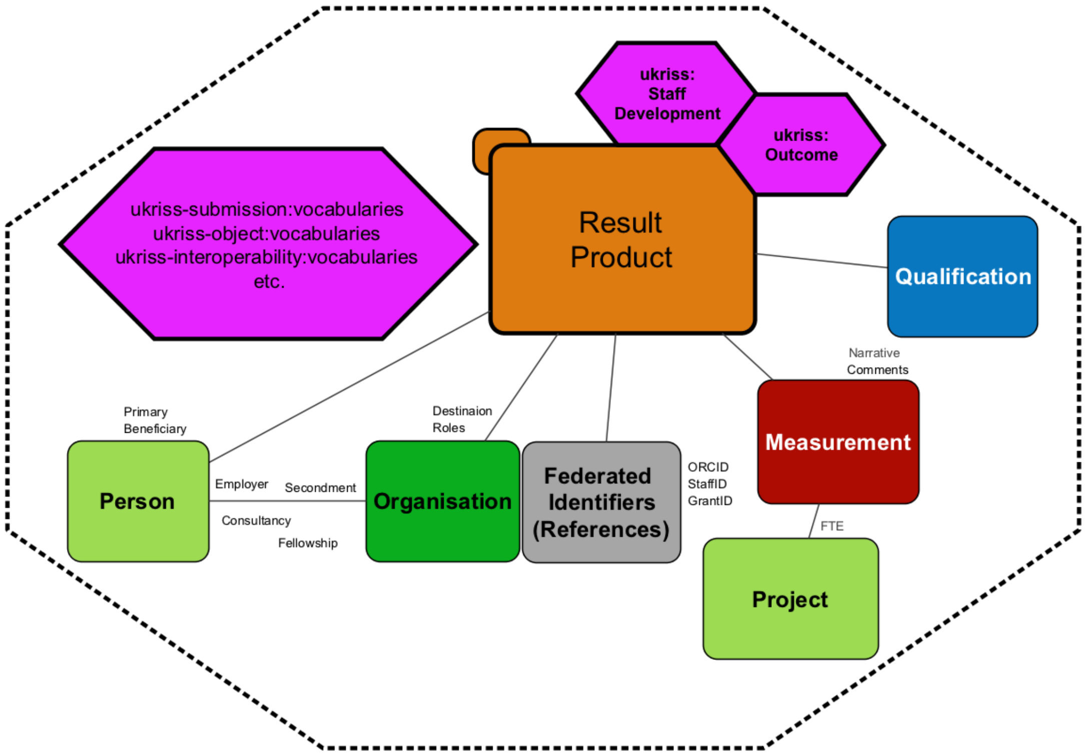


Figure 11.6: Aggregation of Staff Development elements through CERIF entities

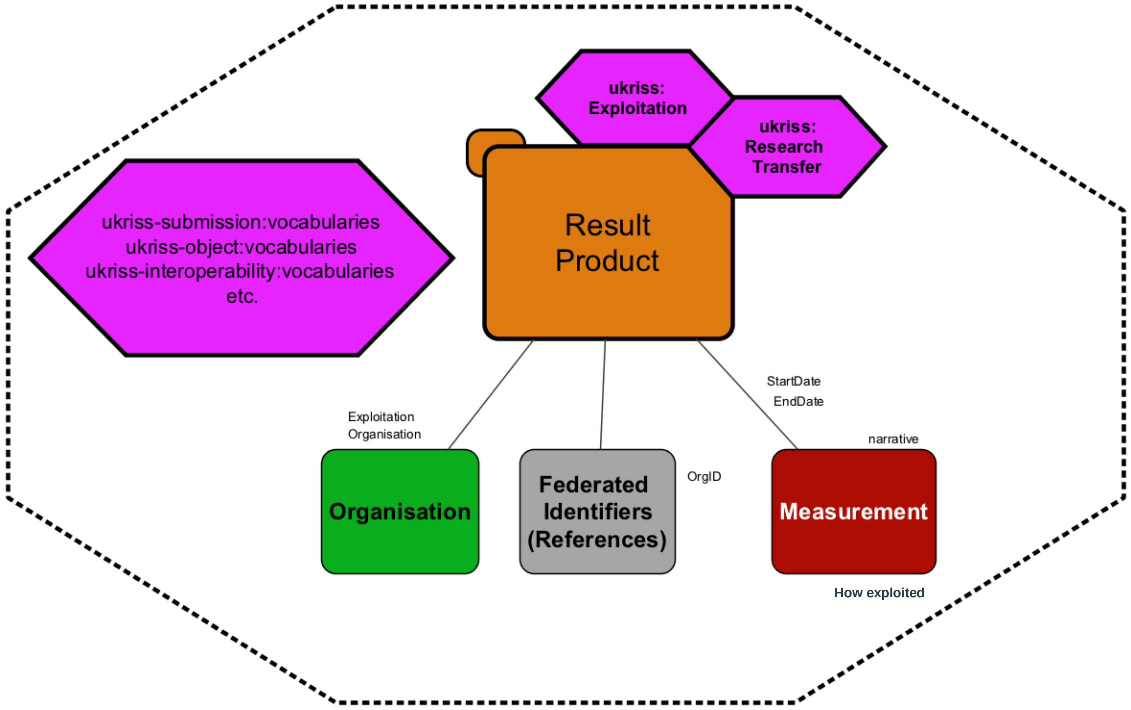


Figure 11.7: Aggregation of Exploitation elements through CERIF entities

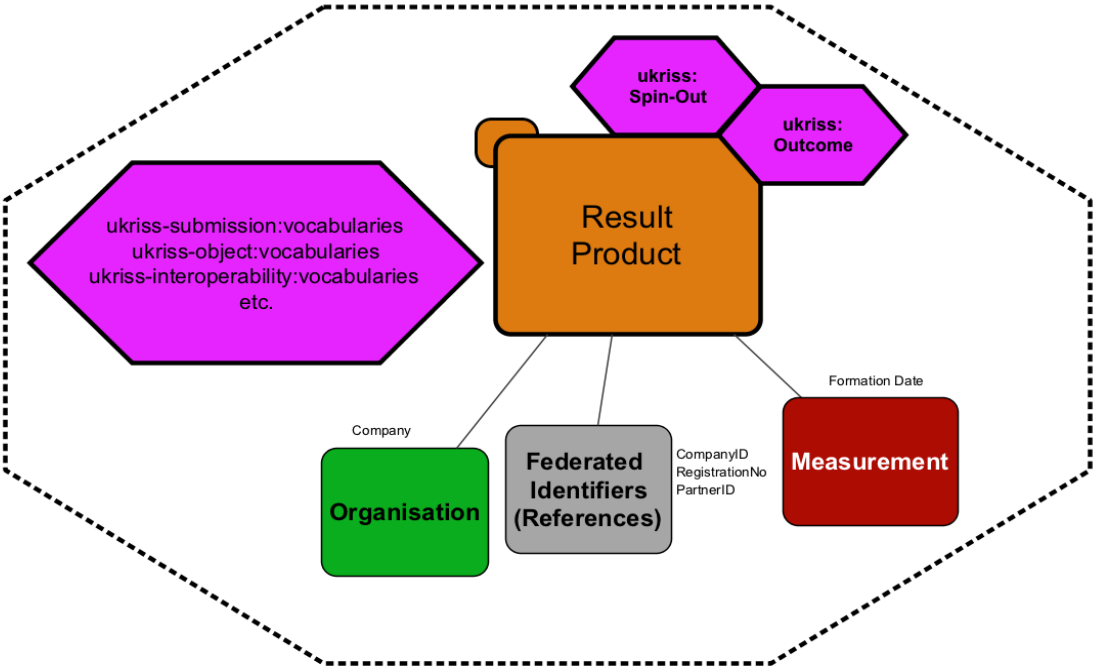


Figure 11.8: Aggregation of Spin-Out elements through CERIF entities

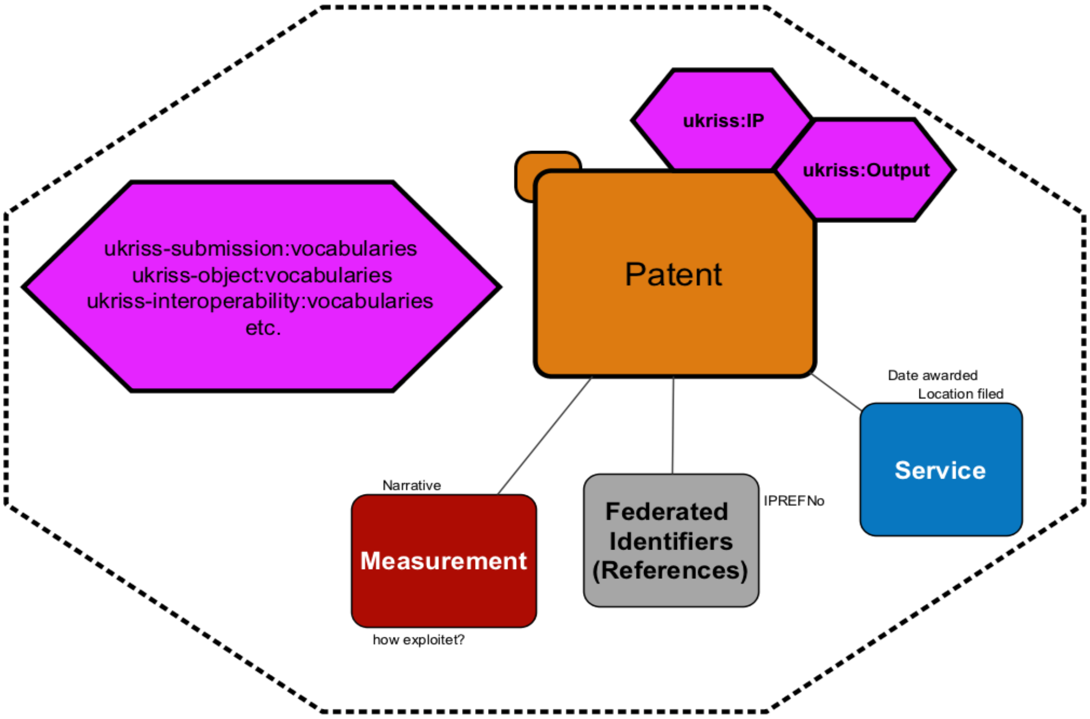


Figure 11.9: Aggregation of Intellectual Property elements through CERIF entities

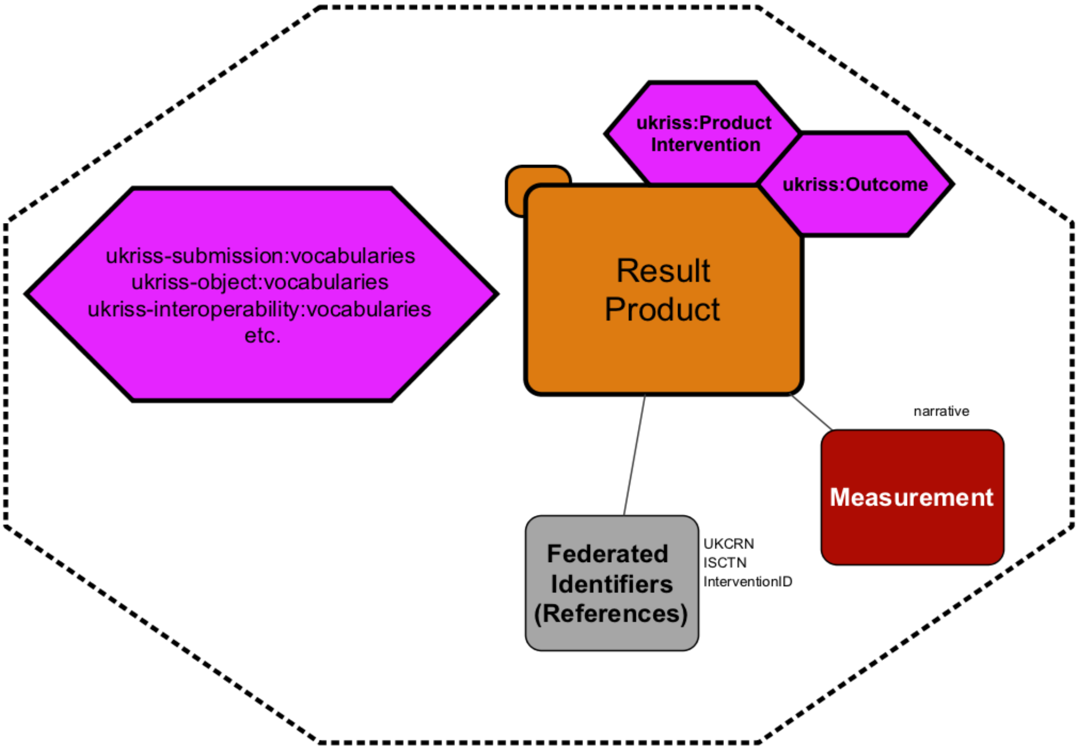

ukriss:Product Intervention
ukriss:Outcome
ukriss-submission:vocabularies
ukriss-object:vocabularies
ukriss-interoperability:vocabularies
etc.
Result Product
narrative
Measurement
Federated Identifiers (References)
UKCRN
ISCTN
InterventionID

Figure 11.10: Aggregation of Product/Intervention elements through CERIF entities

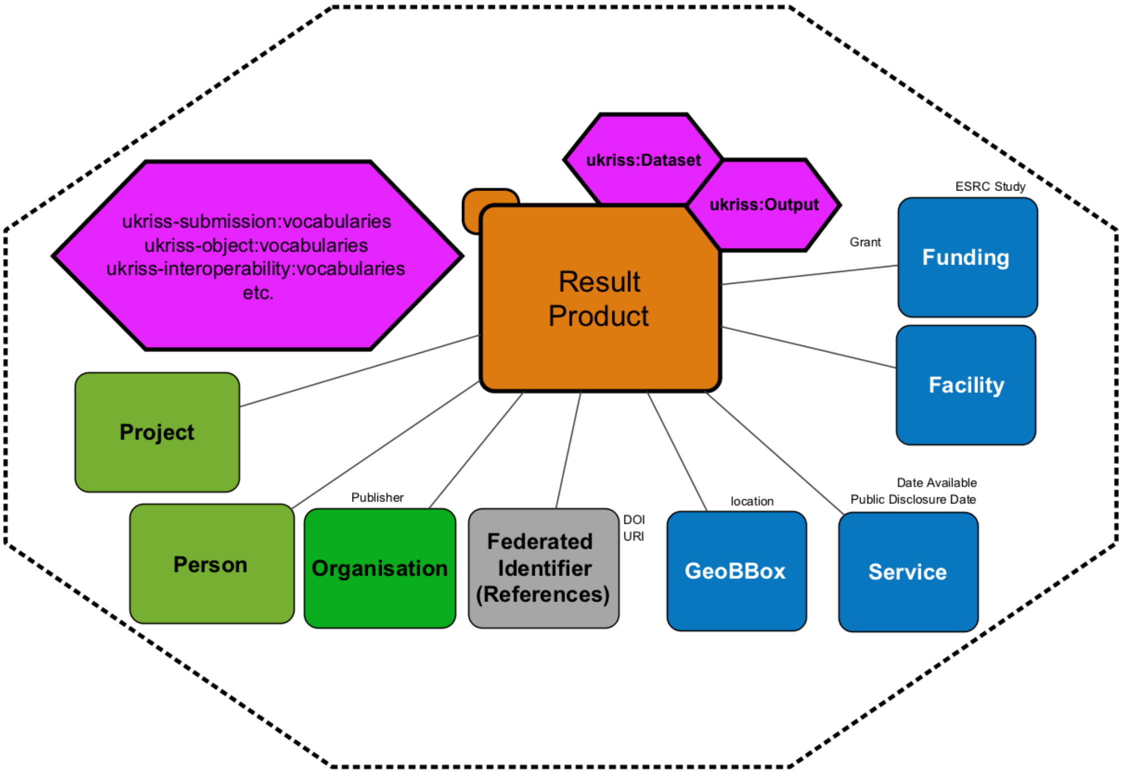


Figure 11.11: Aggregation of Dataset elements through CERIF entities

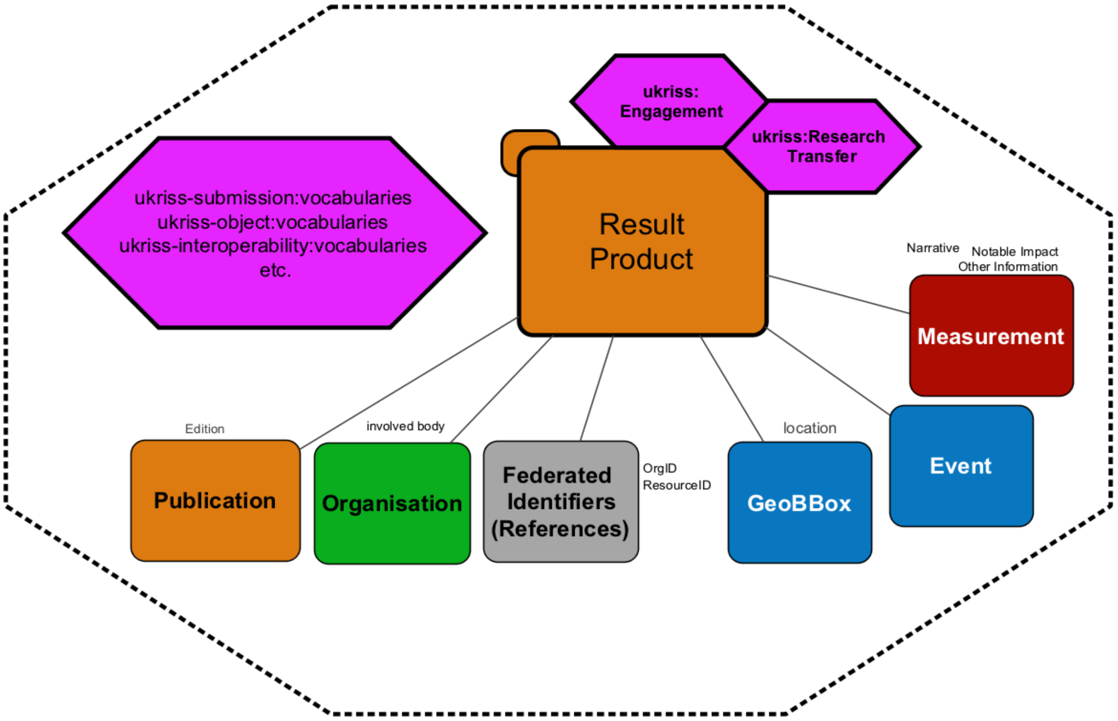


Figure 11.12: Aggregation of Engagement elements through CERIF entities

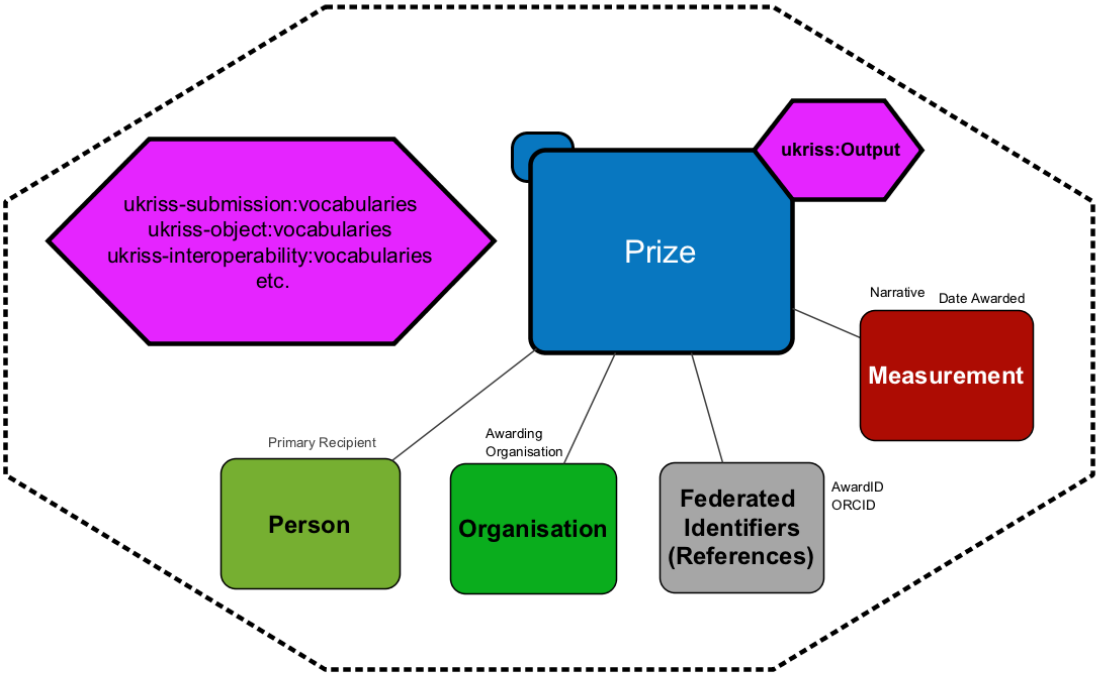

ukriss:Output
ukriss-submission:vocabularies
ukriss-object:vocabularies
ukriss-interoperability:vocabularies
etc.
Prize
Narrative
Date Awarded
Measurement
Primary Recipient
Awarding Organisation
Person
Organisation
Federated Identifiers (References)
AwardID
ORCID

Figure 11.13: Aggregation of Award / Recognition, Prize elements through CERIF entities

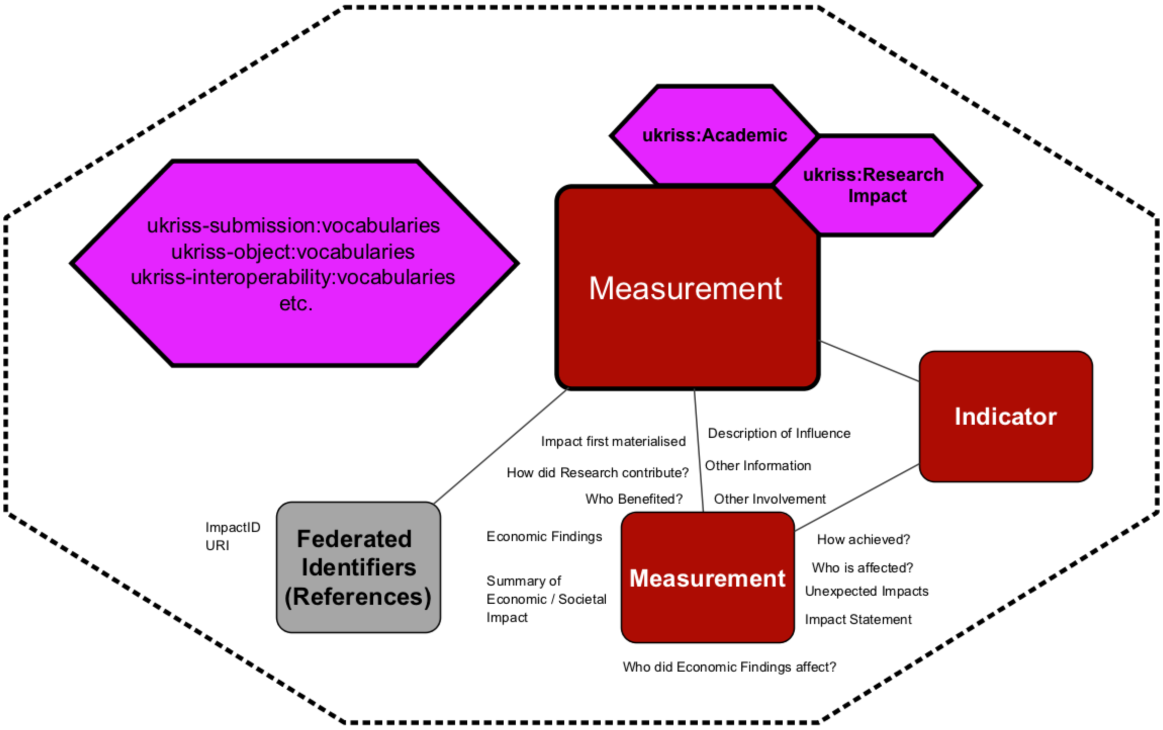

ukriss-submission:vocabularies
ukriss-object:vocabularies
ukriss-interoperability:vocabularies
etc.
ukriss:Academic
ukriss:Research Impact
Measurement
Indicator
Impact first materialised
How did Research contribute?
Who Benefited?
Description of Influence
Other Information
Other Involvement
ImpactID
URI
Federated Identifiers (References)
Economic Findings
Summary of Economic / Societal Impact
Measurement
How achieved?
Who is affected?
Unexpected Impacts
Impact Statement
Who did Economic Findings affect?

Fig 11.14: Aggregation of Impact elements through CERIF entities

12. Appendix 4: Core Profile spreadsheet

The full details of the UKRISS Core Profile and vocabularies are contained in a separate spreadsheet appended to the report.

13. Appendix 5: Visualisation, validation and aggregation

This Appendix details the technical work carried out during phase 2 of the project. The work is divided into three broad categories:

- **Validation** – the process of ensuring compliance with the model and the quality of the data represented
- **Visualisation** – the process of turning the model into a visual representation
- **Aggregation** – the process of bringing together and exploring many objects which are all expressed using the model

13.1 Validation

The validation work was subdivided into two main areas:

1 The process of validation itself.

2 The creation of an example dataset against which content could be validated.

We created a validation framework called “metatool” and a dataset we called “The Academic Catalogue” (ACAT) and used them in tandem to show both how validation could proceed and how external data-sources could be used to cross-check metadata records with each other.

13.1.1 The process of validation

Validating a metadata document is not just simply confirming that the document complies with the rules and specifications of the format it is expressed in, but about ensuring that the content of the document itself is valid. There are three tiers of validation that we looked at in this project, each building on the previous one:

1 **Format Validation** – check that each field conforms to the intended schema or form. For example, we could check that an ISSN is of the form *nnnn-nnnn*, and then we could go on to check that its checksum digit is legitimate. This gives us a high confidence in the validity of the ISSN, but does not tell us whether the ISSN is actually real.

2 **Reality Validation** – see if we can locate the value of the field in some remote dataset, to increase our confidence that it is genuine, and not just a good-looking fake. For example, if we have a DOI which fits with the overall structure of a DOI (for example, by applying a regular expression), we might then go and look up the DOI in the CrossRef database, or follow it to the digital object that it identifies, in order to check that it really exists in the real world.

3 **Cross-Reference** – once we have validated all of the individual fields, we may have retrieved from external sources (via (2)) a number of other records which may include interesting information about the document we are attempting to validate. So, we compare all fields from both the document and the external data, and attempt to determine if the sets of fields are coherent as a whole. For example, if we request information about a DOI from CrossRef, then we also get back a bibliographic metadata record, which we can compare with the one we are validating.

In order to apply these stages to the UKRISS model, we built a framework called metatool, which takes the approach to validation as shown in figure 13.1.

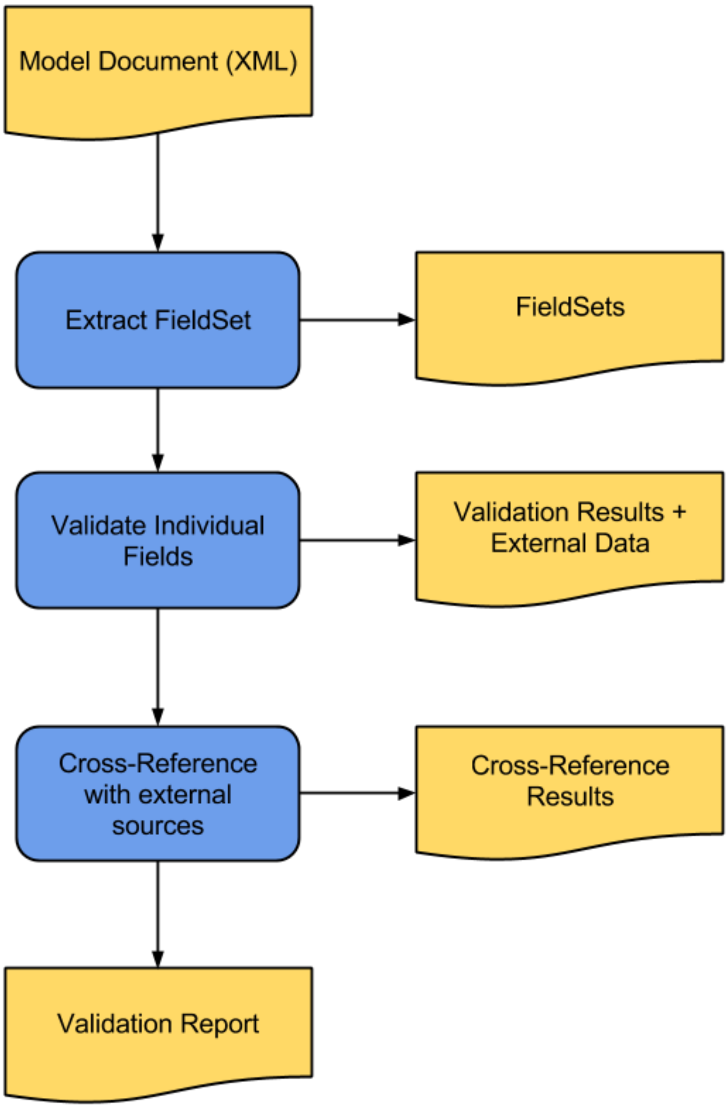

Model Document (XML)
Extract FieldSet
FieldSets
Validate Individual Fields
Validation Results + External Data
Cross-Reference with external sources
Cross-Reference Results
Validation Report

Figure 13.1: Metatool validation pipeline

The primary outputs of the UKRISS phase 2 are models for a variety of aspects of the research information space, including Research Outputs (eg Journal Publications), Collaborations and Partnerships, Exploitation/Spin-Outs, and Engagement Activities. Each of these will be expressed as CERIF, which means that there is a corresponding XML serialisation which represents such artefacts. In figure 13.1 this is the Model Document (XML) which is the input into metatool – an XML record representing the bibliographic metadata and other relationships of a journal article, for instance.

We extract from that XML what metatool refers to as "FieldSets", which are coherent collections of metadata about a research artefact. A single CERIF XML document may contain several FieldSets about different aspects of the artefact. For example, a journal article will have a set of bibliographic information, but also may have information about organisations who hosted the researchers who carried out the research. These may be represented as different FieldSets, so metatool can validate the bibliographic metadata independently of the organisational metadata.

For each field in the FieldSet we can then carry out the validation just based on the expected data-type (eg a field is expected to contain an ISSN) of the field and its value (tier (1) and (2) in the validation process described above). This involves applying a custom piece of code for each data-type, and allowing it to examine first the form of the field, and then to explore external data sources to determine its reality. If, in the process of exploring that external data, the system comes upon additional metadata, it will store it for use in cross-referencing.

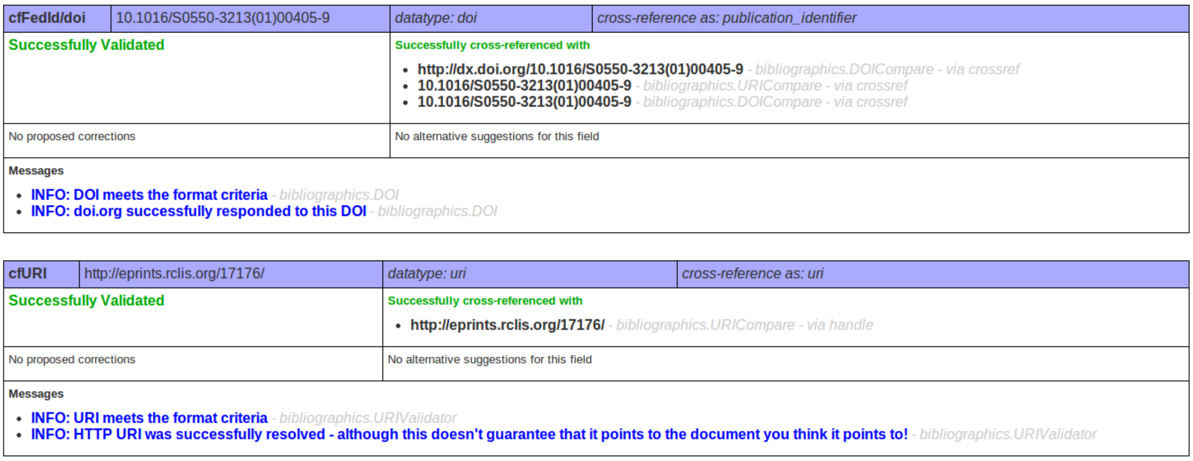
Finally, then we cross-reference the whole record against any external data-sources (tier (3) in the validation process). This involves looking at each field in the data being validated, and comparing it to all equivalent fields in the external data. So, for example, take the title field from the source data and compare it to the title field from CrossRef (2013) and Entrez and determine whether the external data sources agree on the exact title that we started from.

The output of the metatool validation process is in two parts:

1 There is a machine-readable and in some circumstances machine-actionable validation report.
2 There is an HTML rendering of the validation report suitable for human end-users.

The report contains any suggested corrections to the data being validated, alternative values that may be equally valid, and information, warnings and errors that occurred during validation. Together these can be used to see the quality of the metadata record and any enhancements that could be made to it.

Figure 13.2 shows an example of the kind of human-readable output that metatool can provide. In this case it has successfully validated a DOI and a URI against external locations.



cfFedId/doi	10.1016/S0550-3213(01)00405-9	*datatype: doi*	*cross-reference as: publication_identifier*
Successfully Validated		**Successfully cross-referenced with** • **http://dx.doi.org/10.1016/S0550-3213(01)00405-9** *- bibliographics.DOICompare - via crossref* • **10.1016/S0550-3213(01)00405-9** *- bibliographics.URICompare - via crossref* • **10.1016/S0550-3213(01)00405-9** *- bibliographics.DOICompare - via crossref*	
No proposed corrections		No alternative suggestions for this field	
Messages • **INFO: DOI meets the format criteria** *- bibliographics.DOI* • **INFO: doi.org successfully responded to this DOI** *- bibliographics.DOI*			

cfURI	http://eprints.rclis.org/17176/	*datatype: uri*	*cross-reference as: uri*
Successfully Validated		**Successfully cross-referenced with** • **http://eprints.rclis.org/17176/** *- bibliographics.URICompare - via handle*	
No proposed corrections		No alternative suggestions for this field	
Messages • **INFO: URI meets the format criteria** *- bibliographics.URIValidator* • **INFO: HTTP URI was successfully resolved - although this doesn't guarantee that it points to the document you think it points to!** *- bibliographics.URIValidator*			

Figure 13.2: An example of the HTML validation output from metatool

Walkthrough

In order to make the process more clear, it is instructive to take a walkthrough of the process on a single metadata record.

Imagine we have a very simple metadata record, which contains only the ISSN and the title of the journal. Without considering the original CERIF XML, assume that we extract from that XML two fields which make up our FieldSet:

ISSN: 1745-6150
Journal: biology direct

First we look at each field independently, and attempt to validate.

1 The ISSN is of the correct form "nnnn-nnnn", and we can confirm that "0" is the correct checksum digit, so this ISSN at least looks realistic.

2 The journal is much more difficult to validate, all we can say at this stage is that it is an alphanumeric string, which is the stuff that journal titles tend to be made of.

Second we attempt to validate in the outside world. In this case we use the ACAT (which is described in more detail later), and we check the ISSN to see if there is any record of it, and if there is any related information. We discover that the ACAT does have information about our ISSN, which looks like this:

```
{
journal_abbreviation: [ "Biol. Direct", "Biol Direct" ],
issn: [ "1745-6150" ],
journal_title: [
        "Biology direct",
        "Biol. Direct",
        "Biol Direct",
        "Biology Direct"
],
electronic_issn: ["1745-6150"],
publisher_name: ["BioMed Central"],
}
```

When we come to do our cross-referencing validation, we compare the journal we started with against the "journal_title" that ACAT is aware of and find that our title is most likely correct, as it is very similar to the ones listed above. We can make the following statements in our final validation report to the end-user:

1 The ISSN is absolutely valid.

2 The journal is valid, but typically "biology" is written "Biology", so we will suggest that the capitalisation is corrected in the record being validated.

There are also some additional things that we could say in the validation report, such as:

1 That the ISSN provided is the Electronic ISSN.

2 That the publisher of the journal is "BioMed Central".

Such information could potentially be used to further enhance metadata records.

13.1.2 Supported fields

Metatool uses a plugin framework which allows support for fields such as titles, ISSNs, journal titles to be added whenever the models or documents that it is validating change or evolve. The following are the field types that we have built plugins for during the UKRISS project; they are principally focused on the publication model, as this is the most clearly understood of the models:

1. CNRI Handle – we can identify when a URL is a handle, and we can verify that a handle is real. It should also be possible to pull some metadata from the handle server to cross-reference against, but this has not yet been done.

2. Language – we can identify languages written using iso-639-1 and iso-639-2 and any languages written in full in English. We can also recommend alternative expressions of a language, such as indicating that "iso-639-1:en" is equivalent to "iso-639-2:eng" and "English".

3. Dates – we can handle a wide variety of date formats, and check that they refer to valid times.

4. Integers – any fields, such as issue or volume number, which should be integers can be easily validated.

5. ORCID – we can identify when a field is an ORCID, and can retrieve information about that identifier from the ORCID service, which can then be used for cross-referencing.

6. ISSN – we can identify the form and calculate the checksum of ISSNs, and can retrieve information about it from the ACAT, which can then be used for cross-referencing.

7. Journal title – we use some simple string analysis to determine whether journal titles are unlikely to be real; it is not a very sophisticated validation.

8. ISBN – we can identify the form and calculate the checksum of ISBNs. We cannot currently retrieve any metadata about an ISBN, but this could be possible from something like Google Books·

9. DOI – we can identify when a URL is a DOI, and we can verify whether it is real. We can also obtain the bibliographic metadata associated with the DOI from CrossRef and use it to cross-reference against other fields.

10. URI – we can validate the form of any field which should be a URI, and in the case where the URI is a URL we can confirm that it resolves to a resource on the web.

11. PubMed ID – we can do basic validation of the form of a PubMed ID (although as a number of between 1 and 8 digits, this is not very sophisticated), but we can look up bibliographic records using the PMID in the Entrez[3] database, and use the data there for cross-referencing.

[3] http://en.wikipedia.org/wiki/Entrez

13.1.3 Comparable fields

Being able to validate a field and being able to compare two fields which should be the same is not quite the same thing. Metatool provides plugins which are capable of comparing two kinds of object to determine whether they are the same or similar. The following fields are supported:

1 ISSN – ISSNs are straightforward to compare, although in some cases they are expressed hyphenated (eg 1234-5678) and in other cases unhyphenated (eg 12345678).

2 Journal title – we use the Levenshtein Distance algorithm[4] to compare journal titles (and any other text fields).

3 DOI – we canonicalise the expression of the DOI into a standard form, and then compare the strings. DOIs can be expressed as strings in a number of different ways (eg info:doi:10.xxxx or http://dx.doi.org/10.xxxx) so it is not sufficient to simply compare them directly.

4 URI – simple string comparison.

5 Page numbers – we may be able to compare start and end pages, or we may need to calculate page ranges during comparison.

6 Title/Abstract – as with Journal Title we use the Levenshtein Distance algorithm with appropriate tolerance of differences to determine whether two titles/abstracts are the same.

7 Dates – we can read a variety of date formats and then determine if they refer to the same point in time.

8 Volume/Issue – any integer fields are trivially comparable.

9 Language – by knowing how to crosswalk between iso-639-1, iso-639-2 and the English expression of a language, we can compare different expressions in different metadata records.

13.1.4 Future work

The metatool framework is extremely extensible, meaning that it is applicable beyond just CERIF and into any other formats that are currently used to represent or interchange metadata. Examples would include Dublin Core, MODS and METS. Given that each of these formats deals with similar or overlapping data (such as article titles), then once the FieldSets have been extracted from those formats, the plugins for validating the individual fields can be reused in each context – there is only the need for one plugin which validates "titles", and data from CERIF, DC, MODS and METS can all be fed through it.

So the future work with metatool could be around simply extending the plugins that it provides, to cover more formats that are of interest to the community, integrating with more external datasets, and extending the coverage of the data-types that it understands. There would also be a lot of value in finding a way for the community to contribute plugins, so organisations can supply their own if they have particular data-types they are interested in.

[4] http://en.wikipedia.org/wiki/Levenshtein_distance

The ultimate goal would be to provide metatool as a service which can be integrated with, for example, an institutional repository's submission workflow such that any metadata is fully validated before making it into the archive. This would allow us to significantly enhance the metadata quality of published artefacts, improve the ability of aggregators to do interesting things with our data, and increase confidence by end-users that they are being given the correct information.

13.1.5 See also

Software on GitHub: https://github.com/CottageLabs/metatool

Blog posts on Metatool:
http://cottagelabs.com/news/ukriss-model-validation
http://cottagelabs.com/news/metatool-making-metadata-better-data

13.2 The Academic Catalogue

As the previous section illustrates, full validation cannot be carried out without external data sources with which to compare the information being validated. The process of validation will need to consume a lot of external metadata.

For bibliographic metadata there are a variety of data sources that are valuable, such as CrossRef and Entrez, each of which houses many millions of metadata records, but which can principally only be discovered via the primary identifiers in use by those systems (DOI and PMID respectively in those cases). We wanted also to be able to demonstrate validation on other fields, and in particular journal-level metadata such as ISSNs, titles and publishers, and this is the role that the Academic Catalogue (ACAT) fulfils, as good, free to access, catalogues of such data do not exist.
The Academic Catalogue provides a simple data structure that allows us to describe the following information about journals:

- ISSNs (both print and electronic)
- Known journal titles
- Known journal abbreviations
- The publisher(s)

We took a bottom-up approach to populating the ACAT, which means that instead of just looking for and aggregating journal-level metadata from other sources, we attempted to distil it from existing bibliographic metadata which referenced journals, ISSN and publishers, since there are vastly more bibliographical data-sources. As a result, the ACAT knows not only the most commonly known titles of journals, but also the variations of that title which appear "in the wild", which is tremendously useful when attempting to link together information across metadata records of varying quality.

To illustrate this approach, consider the following basic DC metadata record for a journal article:

title: Frogs and Frog DNA
author: Jones, Richard

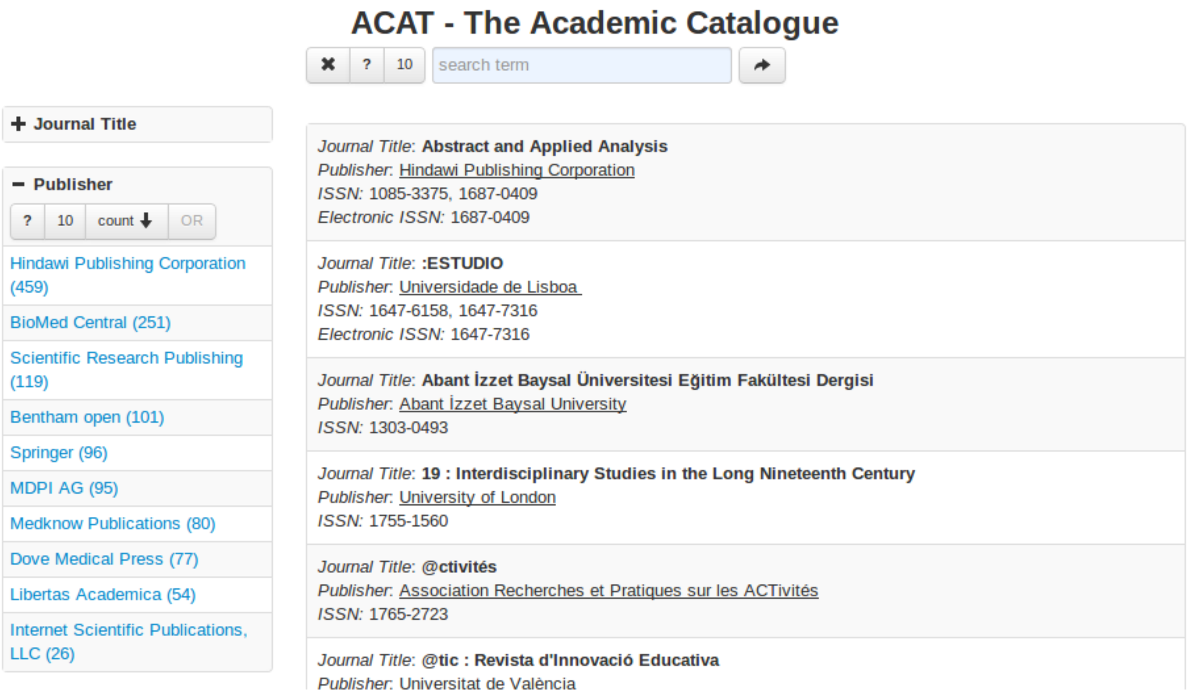
journal: Biology Direct
issn: 1745-6150
publisher: BMC

There are three pieces of metadata here that we are interested in: journal, ISSN and publisher. Since they have been seen together in a single bibliographic record, we make the assumption that they refer to the same thing, so that the journal from BMC called Biology Direct has the ISSN 1745-6150. We can therefore either:

A. Add this as a record to the ACAT.
B. Update any existing record in the ACAT which has either the same publisher, journal title or ISSN, with any additional information from here that it does not already have.

In the previous section we had an example of a record in ACAT for this exact journal, but in that case the publisher was "BioMed Central" rather than "BMC" as we see here. Therefore, we would not add a new record to ACAT in this case, rather add "BMC" to the list of known names of the publisher of the record.

This brings up the issue of provenance, and how reliable are the values in the ACAT. After all, if the source metadata record is wrong, then the ACAT will be wrong, and since we are concerned with validation, this problem needs to be addressed. As part of developing ACAT we looked into provenance, and have developed plans (but no implementation at this stage) to record the source of each part of the journal-level metadata. The more sources we see a particular value come from (eg "BioMed Central" as the publisher, rather than "BMC") the more confidence we can have that that is the "correct" or at least "canonical" value for that field. In addition, we could rate the sources based on reliability; if we see particular values coming from the Entrez database, we might trust it to be correct over different values coming from a single institutional repository, for example.



ACAT - The Academic Catalogue

+ Journal Title

− Publisher

Hindawi Publishing Corporation (459)
BioMed Central (251)
Scientific Research Publishing (119)
Bentham open (101)
Springer (96)
MDPI AG (95)
Medknow Publications (80)
Dove Medical Press (77)
Libertas Academica (54)
Internet Scientific Publications, LLC (26)

Journal Title: **Abstract and Applied Analysis**
Publisher: Hindawi Publishing Corporation
ISSN: 1085-3375, 1687-0409
Electronic ISSN: 1687-0409

Journal Title: **:ESTUDIO**
Publisher: Universidade de Lisboa
ISSN: 1647-6158, 1647-7316
Electronic ISSN: 1647-7316

Journal Title: **Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi**
Publisher: Abant İzzet Baysal University
ISSN: 1303-0493

Journal Title: **19 : Interdisciplinary Studies in the Long Nineteenth Century**
Publisher: University of London
ISSN: 1755-1560

Journal Title: **@ctivités**
Publisher: Association Recherches et Pratiques sur les ACTivités
ISSN: 1765-2723

Journal Title: **@tic : Revista d'Innovació Educativa**
Publisher: Universitat de València

Figure 13.3: A basic interface over the Academic Catalogue, showing the breakdown of records by publisher on the left

13.2.1 Data sources

During the UKRISS project we were only able to investigate two data-sources to populate the ACAT. One was a bibliographic database and the other a journal-level database:

1. MedLine – this is a large dataset which is a subset of the NCBI Entrez database, and which can be obtained in XML after registering with NCBI (for free). It contains 25 million records from the biosciences, and by mining a small portion of it we extracted 10,000 highly consistent journal records.
2. Directory of Open Access Journals – demonstrating that we could obtain data from both bibliographic records and journal records, we incorporated the DOAJ dataset into ACAT, which was very straightforward with their CSV download option.

We also investigated several other potential sources of data, though, that we would anticipate any future ACAT work would incorporate:

1. SHERPA: RoMEO, JULIET, FACT – databases of publisher policies, which relate ISSNs, publishers and journal titles.
2. Mendeley – a large, social, reference management and research sharing tool, which contains a large corpus of bibliographic records.
3. ORCID – researcher identifier system, which also lists publications associated with an author.
4. British National Bibliography – The British Library's bibliography of 3 million records.
5. KB+ – Jisc knowledge base for e-resources.
6. Harvesting from all subject/institutional repositories.
7. Global Open Knowledge Base.

With the number of bibliographic records and journal-level metadata records, we believe that near-full coverage of ISSNs would be possible. Our preliminary work using just a small section of MedLine and the DOAJ gave us 20,000 records, which is just the tip of the iceberg.

13.2.2 Future work

The ACAT has the potential to be a large and comprehensive database providing authority information in a variety of contexts. There has been interest during demonstrations of ACAT in adding more metadata at the journal level, such as legitimate volume and issue numbers, as well as indications that other high-level entities such as conferences could be valuable to represent.

The work so far demonstrates that constructing such a catalogue is possible, and that the data sources available to populate it exist and are accessible. There are some questions to answer over re-licensing of the data, but given the factual nature of much – if not all – of the metadata, we will only need to be concerned with the database rights of the sources.

The main work for the future, for ACAT, then is to incorporate more sources and to actually build the comprehensive dataset, while at the same time incorporating the provenance/audit trail required to build trust in the content.

13.2.3 See also

Catflap, software which provides access to ACAT: https://github.com/CottageLabs/catflap

Blog posts on ACAT and DOAJ:
http://cottagelabs.com/news/using-doaj-data
http://cottagelabs.com/news/an-academic-catalogue

13.3 Visualisation

The objective of the visualisation work was to provide an interactive web-based visualisation of the models, which would allow both exemplar model documents to be visualised to document the project outputs, but also to visualise real-world serialisations of data from institutions using the model. This work could therefore benefit not only the project itself but also users of CERIF in general.

The first stage of developing the visualisation was to determine what general form it would take. We reviewed a number of standard visualisations including:

1. Force-directed graphs – where nodes and edges arrange themselves based on the strength of their connections to each other.
2. Circular packed layouts – where a hierarchy of nested circles shows containment relationships.
3. Tree-node diagrams – where tree-like information structures are represented expanding from some root node.
4. Indented-tree diagrams – where tree-like information structures are represented as like a list structure, with indents to show nesting.

The challenge was to determine which kind of diagram would be most appropriate to represent the graph-like nature of the relationships modelled in CERIF. Therefore, force-directed graphs initially seemed like the most appropriate route to go down. It became quickly apparent, though, that this kind of diagram would be largely unusable as documentation, due to its non-deterministic output (the graphs do not always look the same, and are hard to predict the shape of in advance), and the quantity of information that we were going to need to present.

We therefore determined that if we could restructure the CERIF model as a tree, then we could employ one of the tree-based layouts. This was relatively easy to do, as serialising CERIF to XML is effectively the process of turning the model into a tree. We could, essentially, simply render the structure of the XML as a diagram, and have our visualisation. It would overlook some of the internal linking between elements in a CERIF XML file, but we would be able to look at ways to overcome that limitation later.

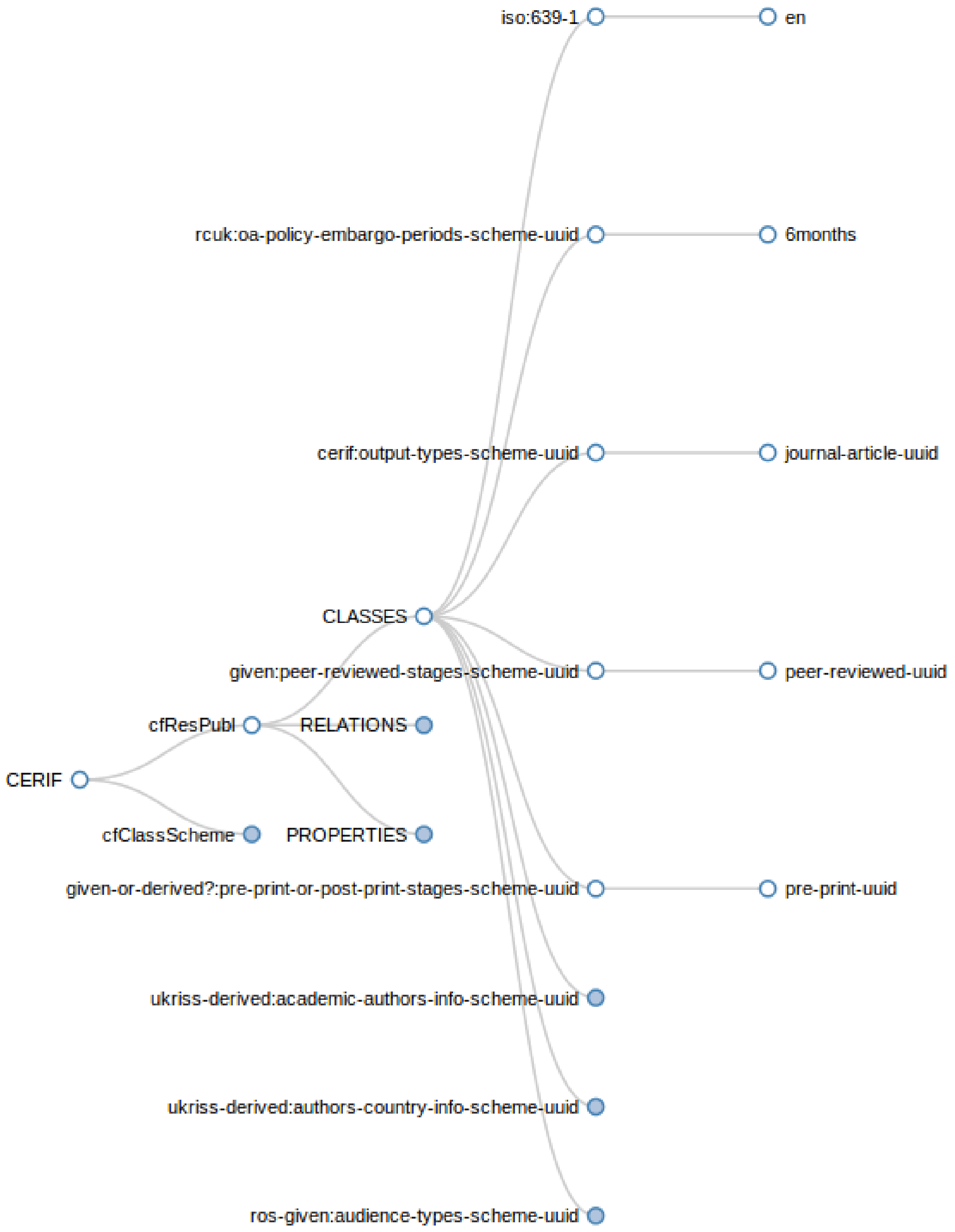

iso:639-1
en
rcuk:oa-policy-embargo-periods-scheme-uuid
6months
cerif:output-types-scheme-uuid
journal-article-uuid
CLASSES
given:peer-reviewed-stages-scheme-uuid
peer-reviewed-uuid
cfResPubl
RELATIONS
CERIF
cfClassScheme
PROPERTIES
given-or-derived?:pre-print-or-post-print-stages-scheme-uuid
pre-print-uuid
ukriss-derived:academic-authors-info-scheme-uuid
ukriss-derived:authors-country-info-scheme-uuid
ros-given:audience-types-scheme-uuid

Figure 13.4: First draft of collapsible tree visualisation

In order to make the diagram more comprehensible to the user, we broke the structure of a CERIF object into its component parts, and separated them in the diagram so they could be viewed independently. The key components of any CERIF object are:

1 The **Classes** applied to the object. These are semantic "tags" which tell us about the nature of the object.

2 The **Relationships** from the object to other objects. These are semantically typed relationships to other objects which may well be in the same CERIF XML file, and therefore may also appear in the visualisation.

3 The **Attributes** of the object. These are the primitive bits of data associated with an object, so in the example of a journal article, the title, volume or issue number would be attributes.

Figure 13.4 shows the first draft of the visualisation where we have simply extracted the tree-like structure from the CERIF XML and grouped the nodes in the tree into either Classes, Relationships or Attributes (or Properties, as we originally labelled them).

A feature of this diagram which is not apparent from the figure is that it is dynamic. Each node can be clicked by the user to either expand or contract the tree beneath it. In this way we attempt to overcome one of the issues which we have with making a good visualisation: how to deal with the sheer size of the data being dealt with. CERIF documents can be very large, contain a lot of objects, attributes, relationships and classes, and displaying them all simultaneously makes the diagrams very large indeed. So, having a dynamic, collapsible tree was an important part of the work.

When we analysed the usefulness of this first draft of the visualisation, there were a number of problems with it:

1 Overall readability – the text is small, the edges thin, and there is a lot of text that needs to be understood by the user for it to make sense.
2 Homogeneity – although we grouped classes, relationships and attributes into their own sections, there was still not a clear enough distinction between them.
3 Detail – a lot of the detail is lost in the conversion to the diagram; for example, some of the lines are relationships which might be typed, while other lines simply indicate that the next node is nested in the node above.

We therefore carried out a second draft of the diagram, which can be seen in figure 13.5.

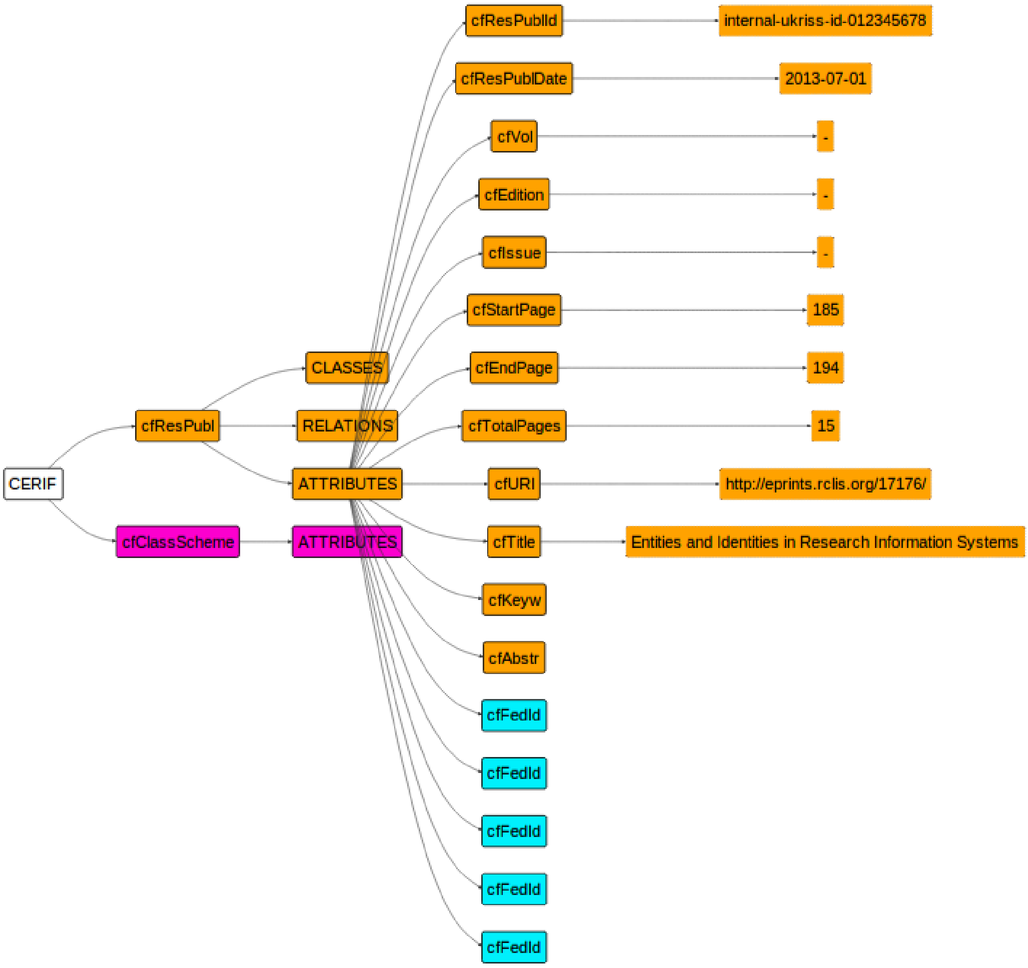


Figure 13.5: Second draft of the visualisation

This tree structure is also collapsible, but we have made a number of enhancements to it. First, the text of the diagram is now more prominent, and forms the basis of each of the nodes, rather than being a label next to the node. Also, the lines which indicate relationships are now labelled with the relationship type. We have also colour coded the various components so that they match with the colours in use in the standard CERIF documentation.

These changes made the visualisation much easier to use, but some of the problems still remained. In particular, the lack of detail; CERIF allows relationships and classes to be annotated with further metadata (such as dates they are applicable) and none of this is easy to include in the diagram.

13.3.1 Future work

There is a long way to go before a good visualisation of a CERIF model or document is available. The models we are working with are large and complex, and they are difficult to visualise as a whole, therefore the visualisation effort will need to be about ways to explore parts of the model in detail while having a view on the overall structure of the model in outline.

The software to work with visual representations of information is also very complex – these basic visualisations took a long time to develop, despite the fact that they look relatively rough. Therefore, any future work in this area will need to take into account the amount of work necessary to make even relatively modest progress.

The likely future of this work is in the development of an exploratory dashboard for CERIF models, which combines both textual and graphical elements. This would allow us to combine a high level graphical outline of the model, for example, and for the user to drill down into that model, and to pull up detailed information in a side-panel that cannot be easily viewed in the main exploration area. Date ranges on relationships would fall into this category, as would the content of large text fields such as abstracts. We would also want to show the relationships between CERIF objects: both those that are within a given CERIF XML document and those which refer to outside entities.

13.4 Aggregation

Aggregation is the analysis and searching of collections of consistently structured data. What this means right now is that it is impossible to aggregate research information, even if it is all pooled together in the same location; this is because the data model and formats for interchange are currently unharmonised.

Figure 13.6 demonstrates a problem that often occurs while trying to perform aggregation or analytics on data coming from data sources which have not agreed on one format.

Dr Jones has only written one article...

Cleaner data reveals Dr Jones has actually been 200% more prolific.

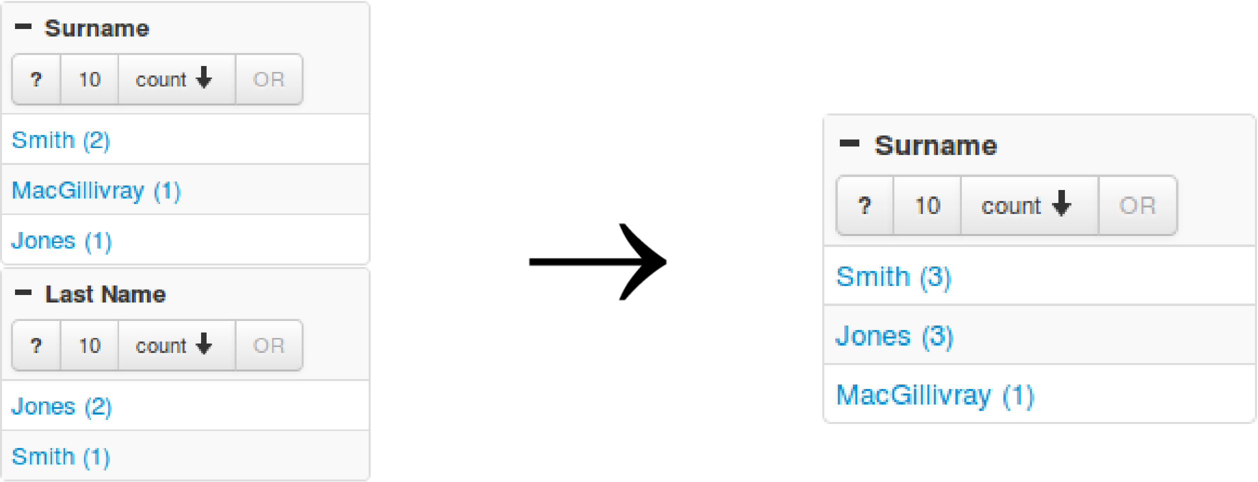


Figure 13.6: The advantages of a consistent underlying data model

As part of the UKRISS project we carried out two exercises around aggregation to demonstrate the value of consistent data modelling:

1 A straightforward aggregation and faceted browse/search interface over a dataset. This allows us to demonstrate that once consistency/harmonisation is achieved, very quickly we can get useful results and interfaces over the data.

2 A more advanced implementation of an aggregation use case (presented in more detail below) showing that once you can do aggregation, you can answer difficult questions of the whole dataset much more easily.

13.4.1 Faceted search/browse

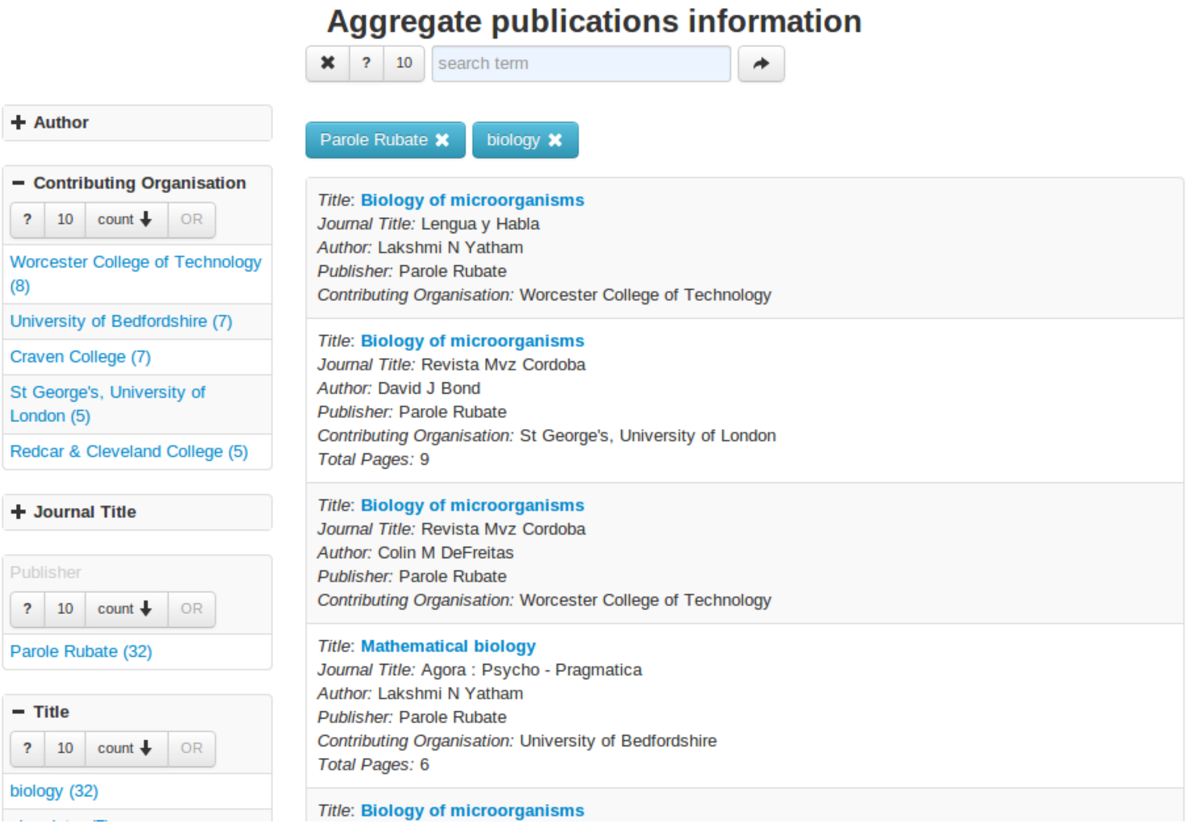
Due to the limited availability during the project of real data from institutions using the UKRISS models, we began the process by generating artificial data of the appropriate structure, so we could demonstrate the value of having such data. We created a search index and view on it using two standard pieces of open source technology:

1 Elasticsearch, an advanced open-source document storage and indexing engine. Elasticsearch can analyse the values stored in the same field across a (potentially very large) number of documents and provide a facet – a count showing how many documents contain the different values.

2 Facetview, a front-end library, which is designed to work with Elasticsearch, providing an intuitive faceted-browse user interface.

Figure 13.7 shows a screenshot of the user interface. Down the left-hand side there are the "facets" – aspects of the data that recur throughout, that we can use to constrain our search results. In the main portion of the window are the results themselves, and at the top is a free-text search area. Using this very simple set-up it is possible to ask relatively complicated questions of the data, such as:

Show me all of the organisations who have publications in journals produced by "Parole Rubate" on the subject of "biology"



Aggregate publications information

+ Author

− Contributing Organisation

Worcester College of Technology (8)
University of Bedfordshire (7)
Craven College (7)
St George's, University of London (5)
Redcar & Cleveland College (5)

+ Journal Title

Publisher

Parole Rubate (32)

− Title

biology (32)

Title: Biology of microorganisms
Journal Title: Lengua y Habla
Author: Lakshmi N Yatham
Publisher: Parole Rubate
Contributing Organisation: Worcester College of Technology

Title: Biology of microorganisms
Journal Title: Revista Mvz Cordoba
Author: David J Bond
Publisher: Parole Rubate
Contributing Organisation: St George's, University of London
Total Pages: 9

Title: Biology of microorganisms
Journal Title: Revista Mvz Cordoba
Author: Colin M DeFreitas
Publisher: Parole Rubate
Contributing Organisation: Worcester College of Technology

Title: Mathematical biology
Journal Title: Agora : Psycho - Pragmatica
Author: Lakshmi N Yatham
Publisher: Parole Rubate
Contributing Organisation: University of Bedfordshire
Total Pages: 6

Title: Biology of microorganisms

Figure 13.7: A simple aggregation faceted browse interface

13.4.2 Advanced use case

As demonstrated by the UKRISS sister project G4HE, there is huge value in providing reports or analytics across aggregate datasets which answer user's specific questions in quite a detailed manner. This is slightly different to providing a flexible faceted browser user interface – the user might not have sufficient information to hand to help them decide how to use it, or they may need to use it in a way that is not immediately obvious to them. Either way, we can provide a custom user interface which answers a user's important question without troubling them with a more complex, flexible system, but which still uses the same underlying data.

The G4HE project carried out an exercise to enumerate some user stories that might be of interest to HEIs, but did not implement them all. Therefore, as part of the UKRISS project, we decided to take one of those user stories and do a demonstration implementation, with the intention of showing that "the coolest thing to do with your data will be thought of by someone else"[5]. We selected the following user story:

As a postgraduate applicant I want to find out which institutions receive the most funding in my subject area because it will help me decide where to apply.

Since we are working primarily with publications data in UKRISS, we altered this slightly to fit our dataset:

As a postgraduate applicant I want to find out which institutions are publishing the most in my subject area because it will help me decide where to apply.

In order to produce a good user experience, this cannot simply be a list of institutions on a page – instead we want the user to receive the best possible information in the most appropriate way. Therefore we decided on the following workflow:

1. The user provides us with some text – for example their master's thesis, or their web page – which covers topics that they are interested in.
2. We text mine the provided text and extract relevant keywords.
3. We search across the aggregation using the keywords to find publications which score highly. We are even able to weight the results such that the keywords appearing in the article title give it more relevance than if they appear in the abstract.
4. We obtain the facet of the organisations whose researchers produced the publications in our result-set.
5. We present to the user an ordered list of organisations, and allow them to view the publications from each organisation that caused it to be brought up as relevant.

The interface developed was not a standard faceted browse interface. Instead, what we have developed is a demonstration as to how services might be created for the benefit of the community, by taking advantage of the power of aggregation.

13.4.3 Future work

Once aggregation is a possibility there is huge potential for future work. The G4HE user stories point to one way forward – the development of services that take advantage of the aggregation to provide tools to specific demographics within the

[5] Rufus Pollock, Jo Walshe, The Many Minds Principle, see: http://en.wikiquote.org/wiki/Free_software

community. There will also be many other user stories which we have not yet thought of. There will also be the potential to join the data in the aggregation up with other datasets to add yet more value.

In relation to the other work in the UKRISS project, it would also be interesting to combine the aggregation and the validation work. While a consistent model would make aggregation possible, there would no doubt still be errors in the data, but by combining the aggregation with metatool we could aim to improve the consistency and therefore the quality of the reporting.

There is no reason why the demonstration user story that we produced during this project could not go on to be a service in its own right, although more work would be required on the business case for such a tool. Similarly for any user story, the primary work for building services would be around whether they have long-term sustainability.

13.4.4 See also

G4HE: http://g4he.wordpress.com
A suite of tools to allow HEIs to perform analytics over data from RCUK.

FundFind: http://fundfind.cottagelabs.com
A service which aggregates funding opportunities for postgraduates and researchers.

14. Appendix 6: Crosswalk Connector

This appendix provides a more detailed description of the UKRISS Crosswalk Connector.

14.1 Connector description

The Connector builds on the successful RMAS connector development, also funded by Jisc, and the technology has been extended to demonstrate the principle of automating the collation of information requested by funders (as far as possible) and automating the transmission of this information.

The connector is based on the open source Pentaho Data Integration Community Edition (PDI CE), sometimes referred to as Kettle (Kettle Extraction Transformation Transport Load Environment). It has a graphical "drag and drop" environment to combine extract, transform and loading steps. Also included is the ability to schedule "jobs" to run at a predetermined time.

Example screenshots of the Connector user interface are shown in figures 14.1(a) and 14.1(b). As shown, the user interface is intuitive to users familiar with a Windows environment. The drag-and-drop features mean that requirements for specialist technical knowledge are minimal.

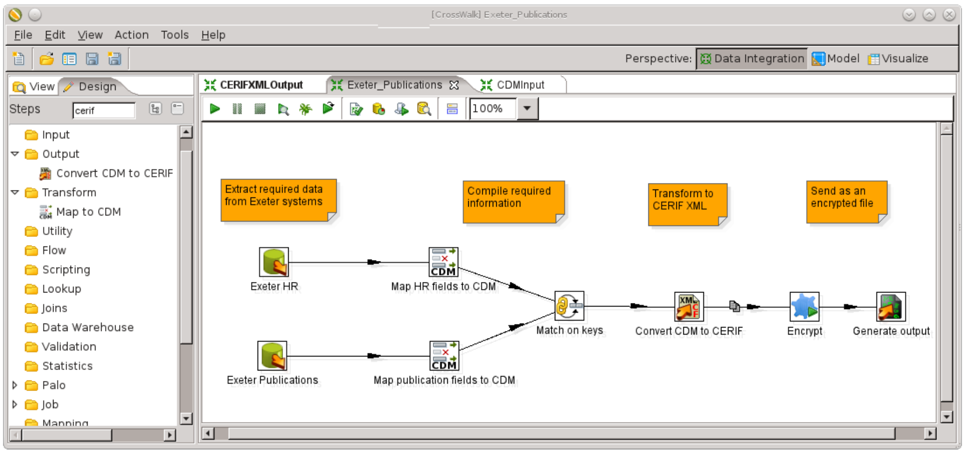
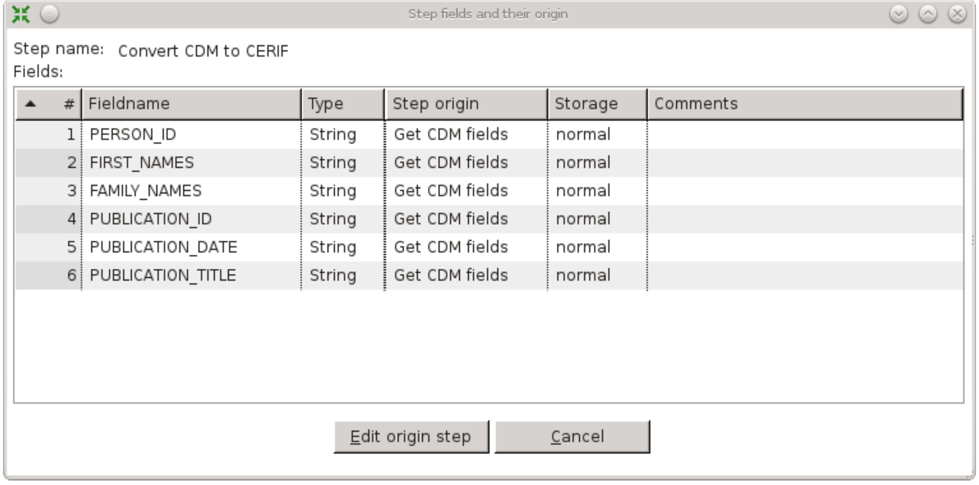


Figure 14.1(a): Screenshot of Connector showing a simple example of a transformto generate CERIF XML.



Step name: Convert CDM to CERIF
Fields:

#	Fieldname	Type	Step origin	Storage	Comments
1	PERSON_ID	String	Get CDM fields	normal	
2	FIRST_NAMES	String	Get CDM fields	normal	
3	FAMILY_NAMES	String	Get CDM fields	normal	
4	PUBLICATION_ID	String	Get CDM fields	normal	
5	PUBLICATION_DATE	String	Get CDM fields	normal	
6	PUBLICATION_TITLE	String	Get CDM fields	normal	

Figure 14.1(b): Screenshot of Connector where fields to be "CERIFed" are specified

14.2 Connector operation

Once the Connector is configured to point towards the correct source servers, such as existing institutional HR, finance systems or project databases, the user interface enables rapid selection of source systems. The user is able to select the fields to be extracted. Simple validation of the data can be applied at this stage to confirm the quality of the data.

Multiple heterogeneous sources of data may be selected and combined. Extraction and load into the Connector is possible for a range of data source formats including XML files, databases and CSV files. This extraction is as a flat data model comprising a set of records each containing a set of data fields, and this is referred to as the Common Data Model. When integrating multiple sources, each containing data relating to the same entity (eg person or project), it is essential to be able to identify matching records. This is achieved in the Connector in a process known as Value Mapping.

In the example of an output to satisfy ROS and Research Fish, the data are then transformed from the Common Data Model into CERIF, using mappings defined by the UKRISS project. For the purposes of the proof-of-concept, publication data was used primarily although the process is not constrained by data type.

Simple error reports are generated in the case of problems with data transfer and mapping processes.

The output from the Connector is CERIF XML, which may be transmitted over an encrypted channel to a destination remote from the institution, and this includes sending directly to the requesting body. Equally, the output may be passed on to additional data validation and aggregation tools (section 3.3.2).

The Connector may of course be used in reverse – funding bodies may use an instance of the Connector to accept and "decode" the CERIF XML, if their internal systems cannot easily consume CERIF XML inputs.

14.3 Example installation at Exeter

By way of simple proof-of-concept in a live environment, the Connector was installed at the University of Exeter. The aim was to evaluate the resource required to install, configure and run the Connector, and also to demonstrate how live raw data may be transformed and transmitted securely in the UKRISS CERIF XML format, with a view to automating the outputs required for ROS and Research Fish reporting.

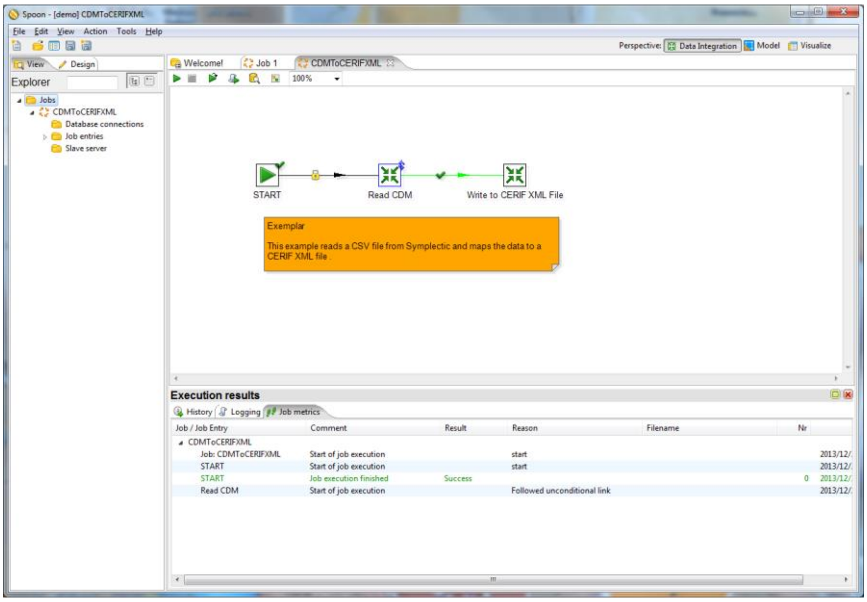
A simple configuration was used, and the resource required for more complex arrangements with multiple inputs and larger volumes of data can be scaled accordingly.

Installation of the Connector, including configuration to accept inputs from the publication management system Symplectic, took one IT technician under one hour.

Once configured, executing a run to generate a CERIF XML output takes seconds, the limiting factor in total time required is how many data quality errors are generated.

More complex installations with several input source systems and outputs containing hundreds of data fields require a little more time to install and configure, perhaps two hours, but the operation time is broadly the same as for simpler configurations. The limiting factor in operation is always the quality of the source data.

Figures 14.2(a)-(c) show example screenshots from the proof-of-concept demonstration, illustrating the drag-and-drop environment and execution results, the input data and the CERIF XML output, respectively.


Spoon - [demo] CDMToCERIFXML
File Edit View Action Tools Help
Perspective: Data Integration Model Visualize
View Design
Explorer
Jobs
CDMToCERIFXML
Database connections
Job entries
Slave server
Welcome! Job 1 CDMToCERIFXML
100%
START
Read CDM
Write to CERIF XML File

Exemplar

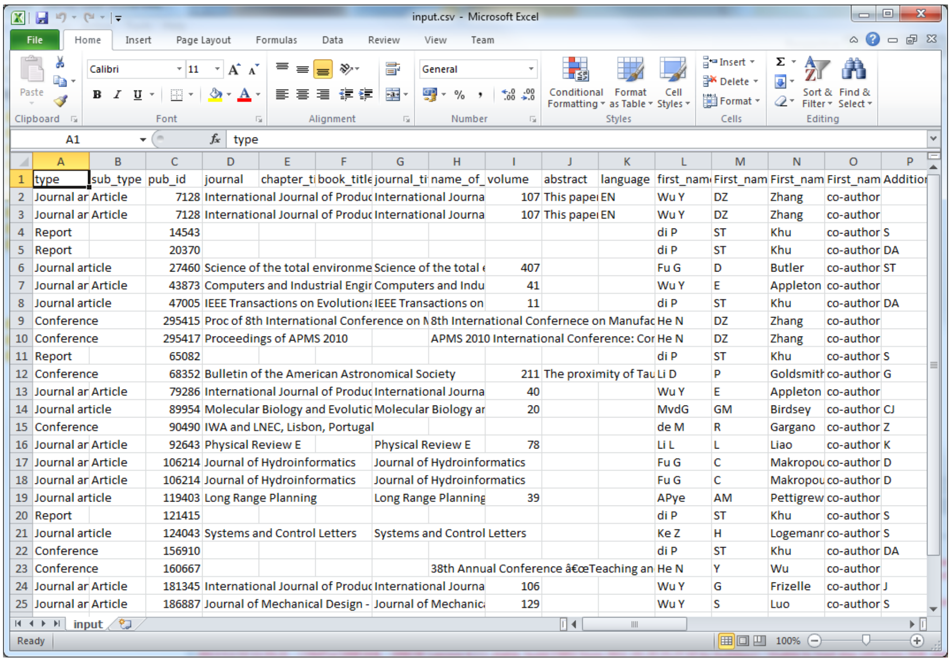
This example reads a CSV file from Symplectic and maps the data to a CERIF XML file .

Execution results

History | Logging | Job metrics

Job / Job Entry	Comment	Result	Reason	Filename	Nr	
CDMToCERIFXML						
Job: CDMToCERIFXML	Start of job execution		start			2013/12/.
START	Start of job execution		start			2013/12/.
START	Job execution finished	Success			0	2013/12/.
Read CDM	Start of job execution		Followed unconditional link			2013/12/.

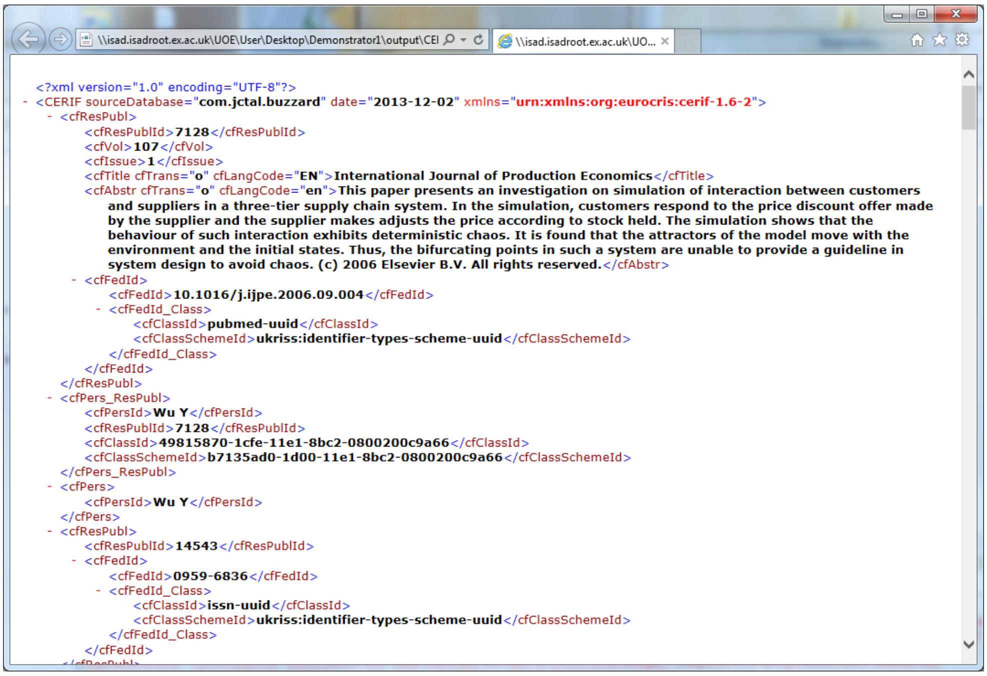
Figure 14.2(a): Screenshot of Connector showing Exeter's simple proof-of-concept configuration (a CSV file input from Symplectic, mapped to create a CERIF XML output), including execution results in the lower panel



f_x type

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	type	sub_type	pub_id	journal	chapter_ti	book_title	journal_ti	name_of_	volume	abstract	language	first_nam	First_nam	First_nam	First_nam	Additior
2	Journal ar	Article	7128	International Journal of Produc			International Journa		107	This pape	EN	Wu Y	DZ	Zhang	co-author	
3	Journal ar	Article	7128	International Journal of Produc			International Journa		107	This pape	EN	Wu Y	DZ	Zhang	co-author	
4	Report		14543									di P	ST	Khu	co-author	S
5	Report		20370									di P	ST	Khu	co-author	DA
6	Journal article		27460	Science of the total environme			Science of the total e		407			Fu G	D	Butler	co-author	ST
7	Journal ar	Article	43873	Computers and Industrial Engir			Computers and Indu		41			Wu Y	E	Appleton	co-author	
8	Journal article		47005	IEEE Transactions on Evolution:			IEEE Transactions on		11			di P	ST	Khu	co-author	DA
9	Conference		295415	Proc of 8th International Conference on N				8th International Confernece on Manufac				He N	DZ	Zhang	co-author	
10	Conference		295417	Proceedings of APMS 2010				APMS 2010 International Conference: Cor				He N	DZ	Zhang	co-author	
11	Report		65082									di P	ST	Khu	co-author	S
12	Conference		68352	Bulletin of the American Astronomical Society					211	The proximity of Tau		Li D	P	Goldsmith	co-author	G
13	Journal ar	Article	79286	International Journal of Produc			International Journa		40			Wu Y	E	Appleton	co-author	
14	Journal article		89954	Molecular Biology and Evolutic			Molecular Biology ar		20			MvdG	GM	Birdsey	co-author	CJ
15	Conference		90490	IWA and LNEC, Lisbon, Portugal								de M	R	Gargano	co-author	Z
16	Journal ar	Article	92643	Physical Review E			Physical Review E		78			Li L	L	Liao	co-author	K
17	Journal ar	Article	106214	Journal of Hydroinformatics			Journal of Hydroinformatics					Fu G	C	Makropou	co-author	D
18	Journal ar	Article	106214	Journal of Hydroinformatics			Journal of Hydroinformatics					Fu G	C	Makropou	co-author	D
19	Journal article		119403	Long Range Planning			Long Range Planning		39			APye	AM	Pettigrew	co-author	
20	Report		121415									di P	ST	Khu	co-author	S
21	Journal article		124043	Systems and Control Letters			Systems and Control Letters					Ke Z	H	Logemanr	co-author	S
22	Conference		156910									di P	ST	Khu	co-author	DA
23	Conference		160667					38th Annual Conference â€œTeaching an				He N	Y	Wu	co-author	
24	Journal ar	Article	181345	International Journal of Produc			International Journa		106			Wu Y	G	Frizelle	co-author	J
25	Journal ar	Article	186887	Journal of Mechanical Design -			Journal of Mechanic:		129			Wu Y	S	Luo	co-author	S

Figure 14.2(b): Screenshot showing part of the input source data


\\isad.isadroot.ex.ac.uk\UOE\User\Desktop\Demonstrator1\output\CEI
\\isad.isadroot.ex.ac.uk\UO...

```
  <?xml version="1.0" encoding="UTF-8"?>
- <CERIF sourceDatabase="com.jctal.buzzard" date="2013-12-02" xmlns="urn:xmlns:org:eurocris:cerif-1.6-2">
  - <cfResPubl>
      <cfResPublId>7128</cfResPublId>
      <cfVol>107</cfVol>
      <cfIssue>1</cfIssue>
      <cfTitle cfTrans="o" cfLangCode="EN">International Journal of Production Economics</cfTitle>
      <cfAbstr cfTrans="o" cfLangCode="en">This paper presents an investigation on simulation of interaction between customers
        and suppliers in a three-tier supply chain system. In the simulation, customers respond to the price discount offer made
        by the supplier and the supplier makes adjusts the price according to stock held. The simulation shows that the
        behaviour of such interaction exhibits deterministic chaos. It is found that the attractors of the model move with the
        environment and the initial states. Thus, the bifurcating points in such a system are unable to provide a guideline in
        system design to avoid chaos. (c) 2006 Elsevier B.V. All rights reserved.</cfAbstr>
    - <cfFedId>
        <cfFedId>10.1016/j.ijpe.2006.09.004</cfFedId>
      - <cfFedId_Class>
          <cfClassId>pubmed-uuid</cfClassId>
          <cfClassSchemeId>ukriss:identifier-types-scheme-uuid</cfClassSchemeId>
        </cfFedId_Class>
      </cfFedId>
    </cfResPubl>
  - <cfPers_ResPubl>
      <cfPersId>Wu Y</cfPersId>
      <cfResPublId>7128</cfResPublId>
      <cfClassId>49815870-1cfe-11e1-8bc2-0800200c9a66</cfClassId>
      <cfClassSchemeId>b7135ad0-1d00-11e1-8bc2-0800200c9a66</cfClassSchemeId>
    </cfPers_ResPubl>
  - <cfPers>
      <cfPersId>Wu Y</cfPersId>
    </cfPers>
  - <cfResPubl>
      <cfResPublId>14543</cfResPublId>
    - <cfFedId>
        <cfFedId>0959-6836</cfFedId>
      - <cfFedId_Class>
          <cfClassId>issn-uuid</cfClassId>
          <cfClassSchemeId>ukriss:identifier-types-scheme-uuid</cfClassSchemeId>
        </cfFedId_Class>
      </cfFedId>
    </cfResPubl>
```

Figure 14.2(c): Screenshot showing part of the CERIF XML output

15. Appendix 7: Preparatory business case

15.1 Introduction

15.1.1 Context

UKRISS phase 2 has been focusing on developing models which will support the reporting of research outcomes. The intention is to enable an open harmonised approach to collection and exchange of research information across higher education in the UK through the development of a core information profile of common fields together with supporting CERIF mappings and dictionaries of terms. The aim is that this will make information exchange more efficient as well as increasing the quality of information on research outcomes. Implementation of a harmonised approach will, however, require change to existing practice within the sector.

The benefits of harmonising the report of research outcomes are already recognised; for example AHRC, BBSRC, EPSRC, ESRC and NERC all require institutions to use the Research Outcomes System (ROS) to report on the research which they fund. Indeed, Research Councils UK (RCUK) is currently reviewing the business needs for reporting across the RCUK councils and the future needs for research outcomes systems, with their findings expected in April 2014.

However, as the feasibility analysis undertaken in UKRISS phase 1 illustrated, institutions as well as non-RCUK funders and other sector organisations could benefit from wider harmonisation of research reporting.

There are three broad options for adoption of harmonisation that could be considered in conducting a full business case:

- A harmonised approach managed by an open standards body on behalf of the UK research community
- A harmonised approach based on an integrated approach adopted by RCUK funders
- The current status quo

While the feasibility study of UKRISS phase 1 identified the opportunities that harmonisation could afford, there were still a number of outstanding questions to be answered before undertaking a detailed business case:

- What risks are involved for organisations adopting harmonisation or not adopting; and how these risks differ for open standards based and non-standards based harmonisation; these risks will vary by type of organisation and size, with the risks and potential impact for non-RCUK funders (eg charities) and smaller institutions being particularly unclear at present
- The full extent of possible indirect benefits and how these might be measured to best convey their value
- What types of costs will be incurred in implementing harmonisation and how these will vary with organisation
- What are the barriers to constructing an effective business case and how these might be overcome

15.1.2 Aims and objectives

The aim of this report is therefore to explore the benefits and risks associated with implementing a harmonisation of research information reporting and data collection across key sections of the public research base, based on open standards and vocabularies. This will help address the current uncertainties and lay the groundwork for construction of an effective business case once RCUK's future plans regarding research reporting are known.

Specific objectives are:

- Describe the benefits (including negative benefits) for each of the stakeholders and how these might be measured
- Describe the risks for each of the stakeholders, their potential impact and how these might be mitigated
- Identify how the costs of implementing harmonisation might be measured
- Summarise the barriers to uptake for each of the stakeholders, with suggestions of how these might be addressed

15.1.3 Approach

As well as drawing on extant work such as the UKRISS phase 1 feasibility study, stakeholder analysis and data gathered from the recent UKRISS phase 2 workshops, we conducted a series of interviews with key stakeholders to explore potential benefits, risks and barriers. The preparatory business case work focused on three key stakeholder groups: the UK research councils (RCUK), charities as well as institutions (research intensive and less research intensive ones). Inclusion of charities was important as it allowed exploration of whether a RCUK integrated solution would also meet the needs of other funders. The number of new interviews (five in total) was limited due to availability of an appropriate representation of stakeholders. The semi-structured interviews covered benefits and dis-benefits, risks, barriers and enablers and costs. Participants were also asked to identify how benefits and costs might be measured within their organisations. To compensate for the limited stakeholder coverage in the interviews, an analysis of previous interviews undertaken as part of the feasibility study was undertaken to try to ensure all stakeholder groups were represented. The analysis consisted of identifying and categorising both direct and indirect benefits and dis-benefits. Benefits realisation mapping was then used to explore further the dependencies and risks to realising the projected benefits.

15.1.4 Layout

This report on benefits, risks and costing of harmonisation of research reporting proceeds as follows. First in section 15.2, the drivers for harmonisation are considered through a SWOT analysis of harmonising research reporting. A brief description of the key stakeholders involved in adoption is also provided. Sections 15.3 to 15.5 discuss the benefits, dis-benefits, barriers and enablers, risks and costs for each key stakeholder in turn. The report concludes in section 15.6, by summarising the key finding followed by the next steps required to develop a business case for harmonisation.

15.2 The drivers for harmonisation of research reporting

Academic research is a fundamental output of UK higher education (HE). Its value lies not only in the development of new knowledge per se; as NESTA highlight, the research base produced by universities plays a central role in the innovation process, producing economic and social benefits. Many stakeholders are involved, including: the government as the funder of research councils, as well as the councils themselves and other funders; research organisations and individuals; others such as vendors of research reporting systems; and business and individuals as the ultimate benefactors of the research. Research reporting is key as it helps manage the return on investment in research as well as disseminating research outputs. Research reporting is, therefore, a fundamental part of the research ecosystem.

15.2.1 The opportunity

The SWOT analysis in figure 15.1 below provides a high-level overview of the strengths, opportunities, weaknesses and threats associated with harmonising research reporting.

Strengths	Opportunities
• Enables better quality, coverage and timeliness of research info • More efficient use of time/resources • Flexible sustainable open framework	• Improve ROI for research • Tracing research impact pathways • Business intelligence for research councils, institutions, researchers, business • Disruptive influence in vendor market • UK can lead in research output, outcomes and impact analysis
Weaknesses	**Threats**
• Benefits depend on sector buy-in • Limited to research-council funded research • Investment required • Change in practice required	• No sector leadership on harmonisation • Perceived lack of competitive advantage • Small institutions may not be able to capitalise on opportunities • Skills gap

Figure 15.1: Swot analysis of harmonising research data

15.2.2 The key stakeholders

Table 15.1 below summarises the interests and influences of the key stakeholders in research reporting and harmonisation. As the analysis illustrates, there are varying levels of interest. In particular, although highly influential stakeholders such as the research councils are highly interested in efficient, quality research reporting, they are potentially less interested in the UKRISS harmonisation solution as they already have partially harmonised solutions.

Typology	Stakeholder	Interest	Influence
Research Funders	Government	High interest in research reporting, especially impact and return on investment; low interest in solution	High
	RCUK and individual research councils	High interest in efficient, high quality research reporting; medium interest in UKRISS solution as already have partially harmonised solutions	High; without their uptake, harmonisation is unlikely to happen

	Charities	High interest in research reporting; medium-high interest in UKRISS solution	Low; needs could potentially be missed
	Other sector funders/agencies (funding councils, Jisc, HESA etc)	High interest in harmonised research reporting; medium interest in UKRISS solution	Medium/High
Research organisations and individuals	Institutions – including research office, researchers	High interest in harmonised research reporting for large research-intensive institutions; low interest for less research intensive institutions	High for research-intensive institutions
Others	Vendors	Medium interest; watching developments but not leading them	Medium – cannot lead uptake but are a key enabler

Table 15.1: Analysis of research reporting stakeholders' interest in harmonisation

Timescales and limited resources meant that it was not possible to explore the benefits and risks for all of the stakeholders. Research councils, charities, institution and vendors were selected based on their influence, interest, potential to be overlooked and enabling capacity.

15.3 Research councils: benefits, risks and costs associated with harmonisation

Research Councils UK (RCUK) manage an investment in research of around £3 billion per annum on behalf of the government. The money is used to fund world-class research, with the UK ranked second to the US in terms of research performance among the G7 economies. The research councils are, therefore, concerned both with the **reporting by institutions** and **reporting to government** on research outcomes as well as the **effective management of the research investment**.

The seven research councils which comprise RCUK already have a range of systems for collecting research-related information from institutions. Research Outcomes System (ROS) is currently used by AHRC, BBSRC, EPSRC, ESRC and NERC. Research Fish is used by MRC and STFC. The councils also report to government and wider society both through a range of annual reports as well as the Gateway to Research (GtR). Je-S is used by all the research councils to support electronic grant applications from their communities.

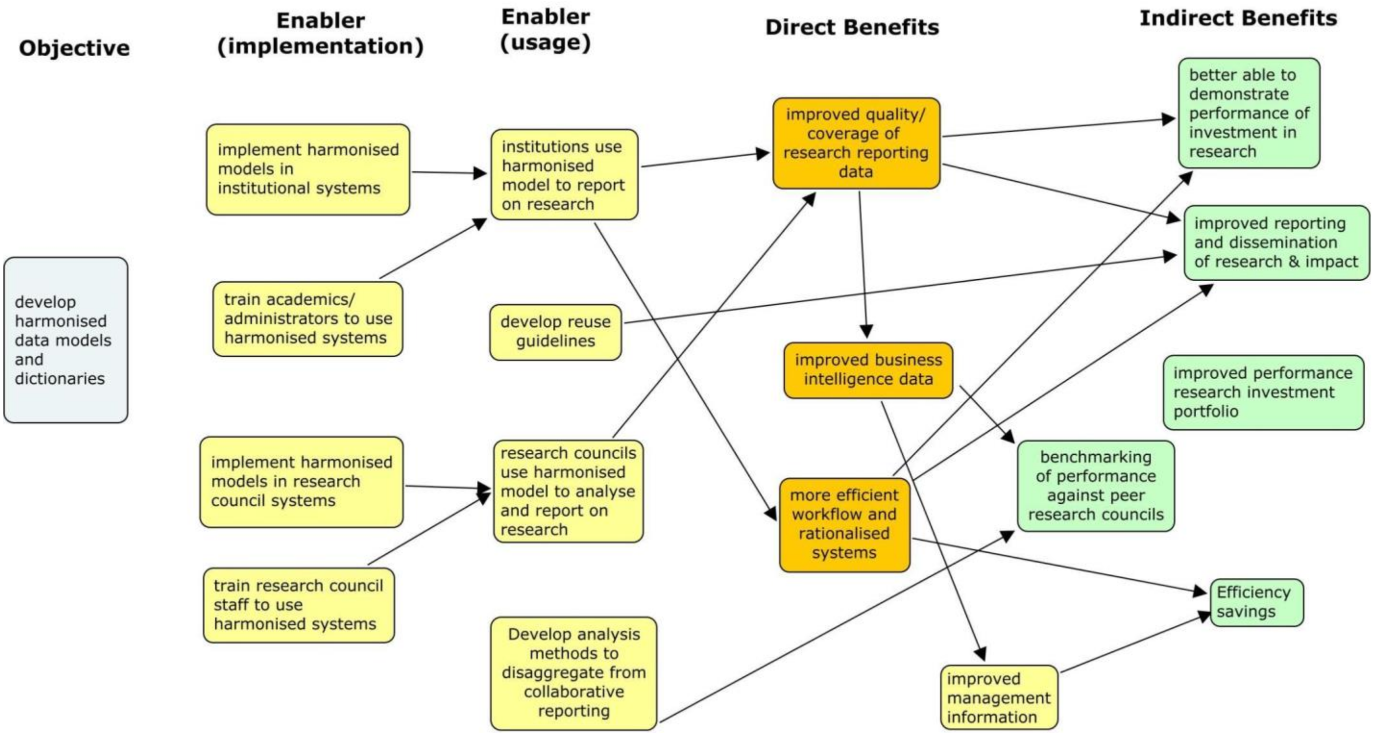

Objective
Enabler (implementation)
Enabler (usage)
Direct Benefits
Indirect Benefits
develop harmonised data models and dictionaries
implement harmonised models in institutional systems
train academics/ administrators to use harmonised systems
implement harmonised models in research council systems
train research council staff to use harmonised systems
institutions use harmonised model to report on research
develop reuse guidelines
research councils use harmonised model to analyse and report on research
Develop analysis methods to disaggregate from collaborative reporting
improved quality/ coverage of research reporting data
improved business intelligence data
more efficient workflow and rationalised systems
improved management information
better able to demonstrate performance of investment in research
improved reporting and dissemination of research & impact
improved performance research investment portfolio
benchmarking of performance against peer research councils
Efficiency savings

Figure 15.2: Outline benefits realisation map for harmonisation of research reporting from a research council perspective

For research councils there are considerable benefits to be gained from harmonisation of research reporting within the UK HE sector. Figure 15.2 above highlights the direct and indirect benefits and dependencies for their realisation.

These benefits and dis-benefits, the barriers and enablers to their realisation and the associated risks for research councils are discussed below. A brief summary of types of costs involved is also provided.

15.3.1 Benefits of harmonisation for research councils

More efficient collection and analysis of data on research

During the current times of economic austerity there has been a considerable drive to ensure maximum returns from money invested in research. Any facility which improves a research council's ability to report accurately on its performance will be beneficial.

The move by some research councils to harmonise reporting through ROS has been driven by the wish to deliver direct efficiency savings as well as improve their individual ability to report on the performance of their investment. Investigation of how to improve efficiency further through wider harmonising of research reporting is ongoing – eg how better organisation of harmonised data and workflows across individual and cross-research council systems could improve the quality of performance management information as well as simplifying processes and provide opportunities for rationalising IT systems. Given the current pressure on research councils' administrative costs, this would deliver significant benefit.

Furthermore, some of the research council information systems have been developed for "tactical purposes" – designed to solve a specific problem. This means that there can be "multiple versions of the truth". A harmonised approach would enable a single source to be developed leading to a more efficient process.

Efficiency benefits could be measured through analysis of the time saved in processing harmonised reporting information.

Better able to demonstrate performance of councils' investments in research, especially contribution to government's impact agenda

The feasibility study undertaken in phase 1 of UKRISS highlighted that harmonisation of research reporting should directly improve the quality of data collected. This in turn would help research institutions to demonstrate the performance of the research investment they are managing on government's behalf.

Furthermore, the data on research currently collected by the research councils do not readily lend themselves to capturing (or illustrating) the full impact of research investment. As one participant recognised, the UKRISS model of how impact is achieved should enable more appropriate data on the impact of research to be gathered and analysed. For example, it will help populate the Gateway to Research (GtR) for disseminating research impact to business and society. This will provide a new window of opportunity, bringing the end-user of research closer to the science, enabling direct feedback to and interaction with the research community. Given the high priority the government has assigned to capturing and illustrating research impact, this was viewed by one research council at least as the primary driver for implementing a UKRISS type approach to harmonisation.

Assessing the economic benefit of being better able to demonstrate contribution to the government's impact agenda is not simple. An increase in the number of different types of impact reported will signal realisation; however, it is difficult to capture its economic value. Illustrative case studies could be used to suggest the scale of economic benefit. For example, the Medical Research Council commented that it was able to trace back new government funding to the evidence it had presented on the research outcomes and return on investment for a previous funding stream.

Benchmarking and improving research investment performance

Better data quality and coverage will enable research councils to improve the analysis of their individual programmes. If the harmonised data are made public or at least shared and across councils that support harmonisation, then cross-funder comparisons can be made, identifying where improvements need to be made and where investment should be focused.

The collective data also provide an opportunity to take a deep dive into the data to try to identify new opportunities to improve the performance of a council's research investment. Such business intelligence capabilities will require some additional investment in systems to be fully exploited.

15.3.2 Dis-benefits of harmonisation for research councils

Ability to reuse data on research reporting inappropriately

Some concern was expressed that data on research could then be reused inappropriately. Two potential issues were identified. First, there was a concern that by harmonising reporting on research, commercial confidential or private data could be exposed. Second, there was a worry that the collected data could be reused out of context. For example, the provenance or source of some data may be lost in the reporting process and so interim findings could be reported as final verified results.

Both these issues could also arise with non-harmonised data; harmonisation simply makes them easier. Good data handling and analysis practices need to be put in place to prevent these risks being realised. For example, Research Fish already has put in place a data release policy.

More difficult to report accurately on individual research council activities and impact

While a harmonised approach to research reporting may enable better capture of impact, especially impact arising from collaborative and/or co-funded programmes, it may negatively impact some reporting. For example, each research council has a responsibility to account for its own spending; however, each council may need to undertake extra work to disaggregate some investments to expose the cross-cutting aspects and ensure that nothing extraneous to their own investment is being picked included in the reporting.

The cost of this extra work should be included in any business case.

15.3.3 Enablers and barriers to adoption of harmonisation by research councils

There is already a degree of harmonisation of information models and dictionaries within the research council as many share the same research reporting system – either ROS or Research Fish – and the benefits in economies of scale can already be seen. To fully implement harmonisation between ROS and Research Fish or to focus on one research reporting system would require extra work and resources.

The economies to be realised through a shared information reporting system for research councils are a key driver for adoption, as is the prospect of being better able to report to government on their return on investment in research. However, successful implementation requires a shift in thinking, especially when dealing with collaborative funding and impact generation. A move from managing information in individual funding programme streams to an ecosystem of information on research is required. A champion would be required to ensure that that information shared through the ecosystem is consistently good and its context fully understood so that invalid analysis is avoided.

15.3.4 Risks relating to harmonisation of research reporting

The benefits realisation map of harmonising research reporting for research councils (figure 15.2) highlights dependencies for the realisation of the anticipated benefits the potential dis-benefits – the risks for research councils in adopting harmonisation.

An analysis of the risks for research councils if they implement harmonisation and if they do NOT as well as to the success of implementation is provided in table 15.2 below. As the analysis illustrates, the risk of not implementing harmonisation is likely to be much more significant than any risks associated with implementing harmonisation.

Risks for research councils in implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Inappropriate analysis of data leads to poor decisions or negative impact	3	2	6	Develop appropriate process safeguards and guidelines which make clear contextual limitations of data; use a champion to drive good practice
More difficult to scope reports to specific investments	2	3	6	Put in place extra processes/systems to facilitate disambiguation
Tied in to a single vendor solution rather than an open solution	4	2	8	A tendering process should ensure that an optimal solution is commissioned; this should include the examination of and safeguards against becoming tied in to a single vendor
Risks for individual research councils in NOT implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Less able to compete for government investment in research	4	5	20	Implement a harmonised approach
Less efficient systems	3	5	15	Implement a harmonised approach
Risks to successful implementation of harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Prohibitive cost	3	2	6	Tendering should ensure value for money; a staged approach to implementation will reduce investment costs initially at least

Table 15.2: Analysis of the risks for research councils associated with harmonisation

15.3.5 Costs for research councils of implementing harmonisation

The costs for implementing harmonisation for the research councils include: software; implementation of new data models or mapping; training; maintenance; and staff time. Additional costs to realise wider indirect benefits are: new hardware; refactoring of legacy data and re-engineering processes; and creation and maintenance of registries.

Information on these costs will be difficult to source for research councils.

15.4 Charities: benefits, risks and costs associated with harmonisation

Charities contribute 5% of the total investment in research within UK. Charity-funded projects range from social improvement and education to medical research. The latter alone contributed over £1.1 bn funding in 2011. While their funding contribution may be smaller compared with that of the research councils, the research funded by charities makes a vital contribution to UK innovation and social well-being. Of prime importance to charities is ensuring that charitable donations are well spent – that they maximise quality research outcomes. Charities, therefore, are concerned with **reporting to the wider research community and society, charity boards, and their contributors** on research outcomes as well as with **reporting by institutions** and the **management of their research investment**.

While some charities use Research Fish, many charities do not have any systems in place for reporting on the research they fund.

For charities, there are some benefits to be gained from harmonisation of research reporting within the UK higher education sector. Figure 15.3 below highlights the direct and indirect benefits and dependencies for their realisation. These benefits and dis-benefits, the barriers and enablers to their realisation and the associated risks for charities are discussed below. A brief summary of types of costs involved is also provided.

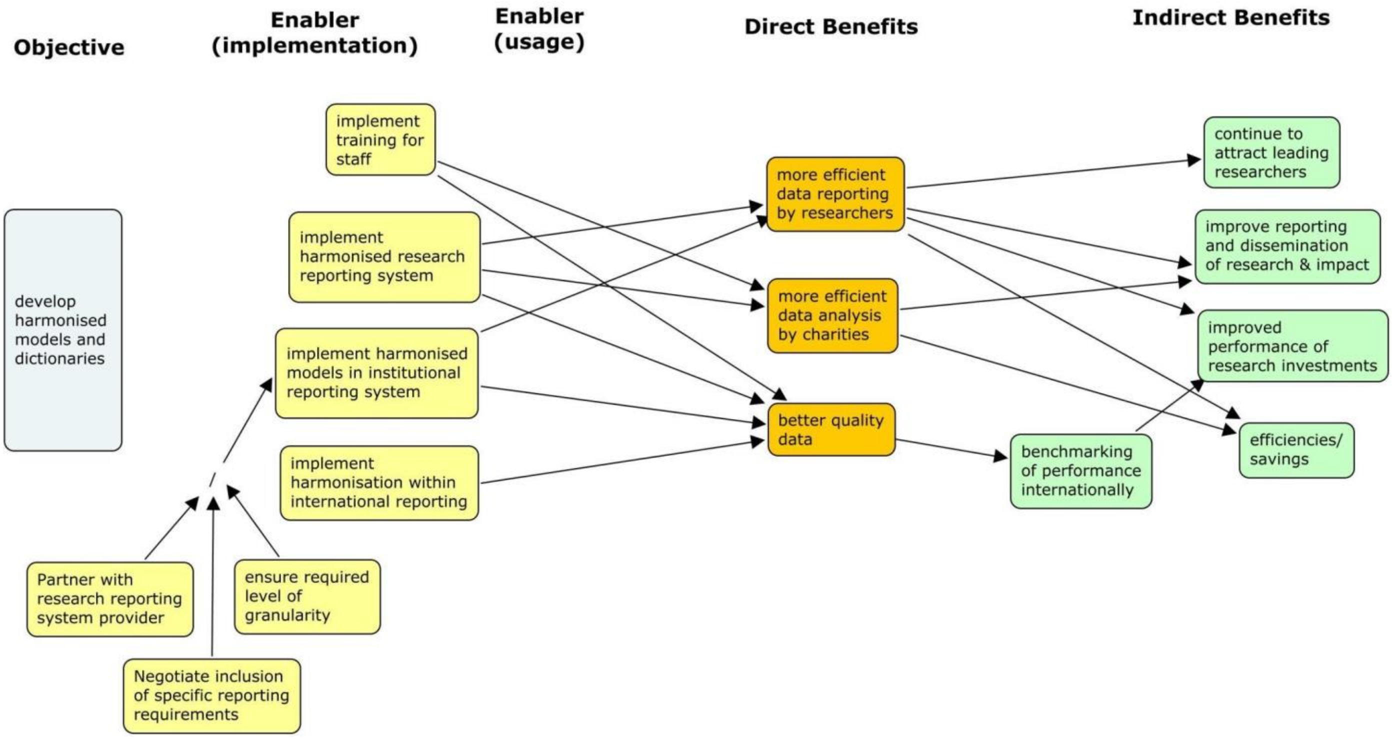

Objective
Enabler (implementation)
Enabler (usage)
Direct Benefits
Indirect Benefits
develop harmonised models and dictionaries
implement training for staff
implement harmonised research reporting system
implement harmonised models in institutional reporting system
implement harmonisation within international reporting
Partner with research reporting system provider
ensure required level of granularity
Negotiate inclusion of specific reporting requirements
more efficient data reporting by researchers
more efficient data analysis by charities
better quality data
benchmarking of performance internationally
continue to attract leading researchers
improve reporting and dissemination of research & impact
improved performance of research investments
efficiencies/ savings

Figure 15.3: Outline benefits realisation map for harmonisation of research reporting from a charity perspective

15.4.1 Benefits of harmonisation for charities

More efficient collection and analysis of data on research

Harmonisation of medical charities' research reporting with the MRC and STFC's Research Fish reporting system has, according to interviewees, already delivered some direct efficiency benefits as well as improved data quality. For example, prior to using Research Fish, some smaller charities did not collect such extensive data and were unable to report on research outcomes as easily. Where research is co-funded between charities and MRC, then use of Research Fish for research reporting may simplify reporting and the sharing of reporting information.

Other charities could derive similar benefits through using Research Fish or ROS; however, the benefits may not be significant enough to warrant the expense for small charities.

Researcher feedback could be monitored for indications that the harmonised reporting was more efficient and internal monitoring could assess efficiencies for charities. Researcher feedback could be monitored for indications that the harmonised reporting was more efficient and internal monitoring could assess efficiencies for charities.

Benchmarking and business intelligence to improve research investment performance

Wider harmonisation of data models and dictionaries with multiple funder or institutional systems is unlikely to deliver any other significant benefits unless the information in these systems is made publically available. If it is, then it might be possible for charities to undertake wider benchmarking of research performance. For example, it could enable tracking of common barriers to commercialisation of research. This insight could then help improve research performance.

It is impossible at this stage to quantify the benefits that could be realised, although the greatest benefits would come through international harmonisation and sharing.

15.4.2 Dis-benefits of harmonisation for charities

A key dis-benefit of harmonisation for charities lies in the fact that due to the potentially prohibitive cost for adopting harmonisation, they may not be able to leverage the advantages that other funders are able to. This could potentially impact charities in two ways. First, if their research reporting requirements are considerably more time intensive for researchers, then researchers – especially those leading their fields – may choose to apply for funding from other sources. This could impact the quality of the research outcome that the charities can fund. Second, the quality of the research reporting data will be much poorer, especially relating to research outcomes. This could impact a charity's ability to convey the usefulness of its research to a wider audience and therefore affect the extent to which the research will benefit the charitable cause it was commissioned to help.

There would also be a dis-benefit if data elements are lost through an aggregation of types in harmonisation. For example, data on fellowships can be held over several data fields (eg clinical, non-clinical) and it would be a risk if aggregation combined multiple fields into one.

15.4.3 Enablers and barriers to adoption of harmonisation by charities

If the harmonisation implemented across the sector is not compatible with their existing research reporting systems then charities are unlikely to adopt harmonised reporting due to the significant costs involved. For the medical charities that means harmonised with Research Fish.

As many charities have no research reporting systems in place, harmonised models and dictionaries could effectively provide a template for implementing a research reporting system. This could be a significant enabler, provided implementation costs were low.

Training demands and lack of technical expertise are also likely to present a barrier to adoption of harmonised systems. Both a support network and a shared research reporting service could aid uptake provided they do not incur significant expense.

Harmonisation with international research reporting systems, for some charities at least, would be a significant incentive for adoption.

15.4.4 Risks for charities relating to harmonisation of research reporting

The charities' benefits realisation map for harmonising research reporting (figure 15.3) highlights dependencies for the realisation of the anticipated benefits the potential dis-benefits – the risks for charities in adopting harmonisation. An analysis of the risks identified through the mapping is provided in table 15.3 below. They are grouped by the risks for charities if they implement harmonisation and if they do NOT as well as to the success of implementation. As the analysis illustrates, the risk of not implementing harmonisation is likely to be much more significant than any risks associated with implementing harmonisation; however, the cost of implementation for small charities means that there is a significant risk that charities may not be able to implement harmonisation of research reporting.

Risks for charities in implementing harmonisation	Impact	Likelihood	Score	Mitigation
Loss of granularity in reporting limits analysis possible	3	2	6	Get involved in development of harmonisation standards to ensure fit for purpose
Harmonisation implemented by service provider does not meet reporting needs	4	2	8	Get involved in development of harmonisation standards and service providers to ensure fit for purpose
Risks for charities in NOT implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Less able to attract leading researchers	4	3	12	Implement a harmonised approach
Less able to achieve desired research outcomes	4	5	20	Implement a harmonised approach
Risks to successful implementation of harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Prohibitive cost	3	4	12	Sign up to a shared research reporting service
Lack of skills to effectively use harmonised approach	2	3	6	Ensure staff are trained; set up support network

Table 15.6: Analysis of the risks for charities associated with harmonisation

15.4.5. Costs for charities of implementing harmonisation

For charities who use services such as Research Fish, it would be expected that any implementation costs would be borne by the service provider; however, there may be an increased service charge for accessing the "harmonised" system. Access to national harmonised data sets for business intelligence analysis would likely incur an additional charge.

Charities implementing their own harmonised solution are likely to incur costs similar to institutions: software; implementation of new data models or mapping; training; maintenance; and staff time. For smaller charities, who aim to spend as much as possible of their donations on research, this could be prohibitive. A collective harmonised solution for small charities may be more affordable.

15.5 Institutions: benefits, risks and costs associated with harmonisation

The degree to which universities and other publicly funded research institutions are involved in the reporting of research information depends on the extent of their research focus. The large, research intensive Russell Group universities receive almost three-quarters of all public research funding. Large/research-intensive institutions are concerned both with **reporting to a range of funders, REF and HESA** on research outcomes as well as with **managing their research profile** and **benchmarking their performance**. The majority of large/research-intensive institutions already have a research information management system, either a CRIS or a repository such as ePrints which has some CRIS capability; although in many cases information on research may be split across multiple systems. Some other universities – often newer and specialist institutions – have much less emphasis on research. These smaller/less research intensive institutions will be less concerned with **reporting to a range funders and the REF.**

For institutions there are considerable benefits to be gained from harmonisation of research reporting within the UK HE sector.

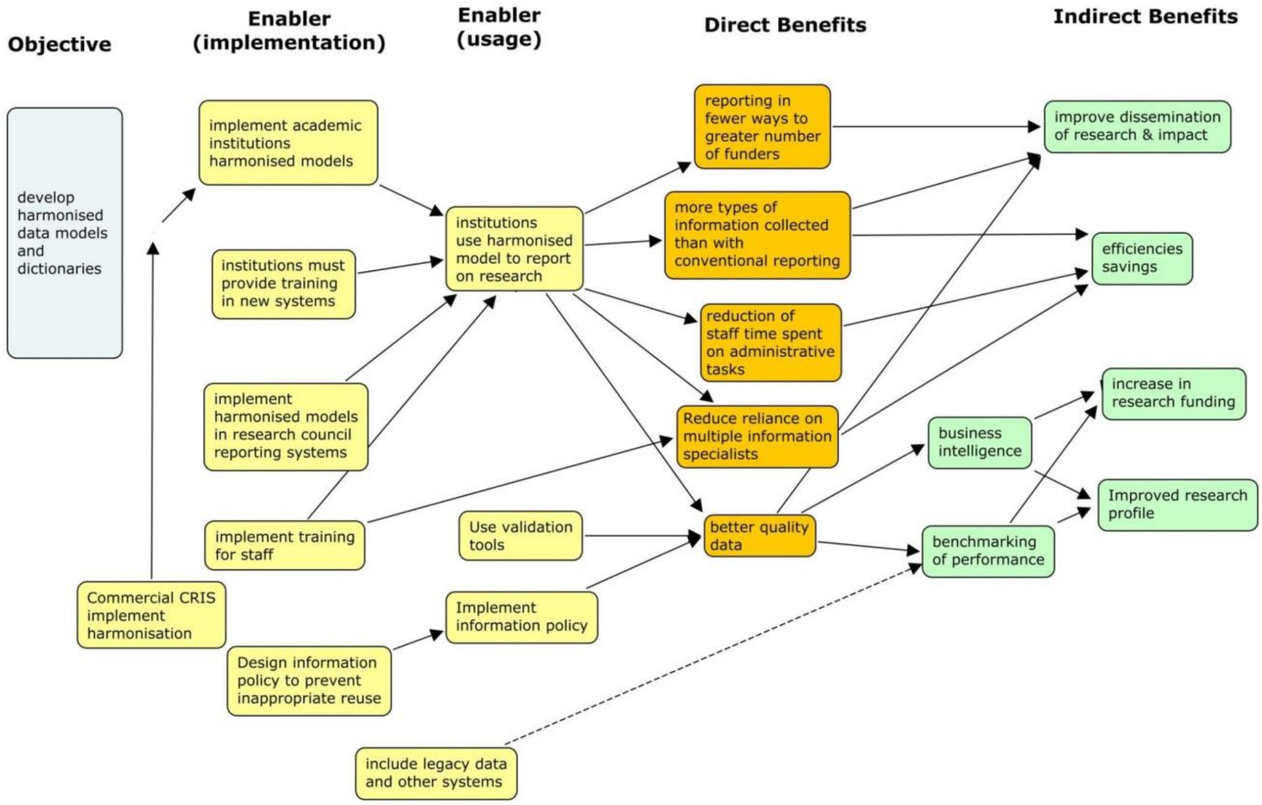

Objective
Enabler (implementation)
Enabler (usage)
Direct Benefits
Indirect Benefits
develop harmonised data models and dictionaries
implement academic institutions harmonised models
institutions must provide training in new systems
implement harmonised models in research council reporting systems
implement training for staff
Commercial CRIS implement harmonisation
Design information policy to prevent inappropriate reuse
include legacy data and other systems
institutions use harmonised model to report on research
Use validation tools
Implement information policy
reporting in fewer ways to greater number of funders
more types of information collected than with conventional reporting
reduction of staff time spent on administrative tasks
Reduce reliance on multiple information specialists
better quality data
business intelligence
benchmarking of performance
improve dissemination of research & impact
efficiencies savings
increase in research funding
Improved research profile

Figure 15.4: Outline benefits realisation map for harmonisation of research reporting from an institutional perspective

These benefits and dis-benefits, the barriers and enablers to their realisation and the associated risks for institutions are discussed below. A brief summary of types of costs involved is also provided.

15.5.1 Benefits of harmonisation for institutions

More efficient collection of better quality data on research

One direct benefit of a harmonised approach to research reporting would be to decrease the time spent by principal investigators or their nominees in completing reports to funders on research outcomes and financial monitoring. Similarly, if harmonisation enabled institutions to have access to the data stored in grant application systems such as Je-S, further efficiencies could be gained. Individual researchers would also directly benefit from more efficient completion of REF submissions, freeing up more time for research.

It is anticipated that harmonisation will also directly improve the quality of the research reporting data through the use of common identifiers (eg people) as well as the proposed UKRISS validation tools. Quality should also be increased, as the reporting system should be less frustrating to use.

The extent of efficiencies could be measured by surveying research staff. Increases in quality would be more difficult to measure, although an increase in data available could be an indicator of at least wider coverage. As harmonisation should actually reduce the data stored, analysis would be required by institutions to identify what actually constitutes additional data coverage. It may be best, therefore, to approach this by exploring indicative pricing of additional information availability using shadow pricing techniques.

Enable management information, benchmarking and business intelligence analysis which were previously infeasible

Harmonisation will bring an institution's information on research together in one coherent environment although not necessarily one system. This will enable types of analysis such as management information, benchmarking or business intelligence which were not previously feasible – due either to the length of time it would have taken or gaps in data. The ability to analyse management information could help identify inefficiencies in an institution's research system – for example by examining success of grant applications, an institution could focus its efforts to reduce time spent in unsuccessful areas or identify how success rates might be increased. If the harmonised information is openly published, benchmarking would allow an institution to compare its performance in key areas of interest with that of comparable institutions. Analysis of business intelligence data could help identify new fruitful areas for focusing research resources or industry partnerships. This analysis could be undertaken either by individual researchers or by Research Services.

Identifying the potential economic value associate with the new analysis capability is problematic, given this has not yet occurred. It may be possible to draw on work in the economic benefits of benchmarking and business intelligence in other contexts to suggest indicative economic benefits.

Remove the need for niche information specialists

Harmonisation will also remove the need for multiple niche information specialists within an institution. While this in itself may not lead to efficiencies as the analysis will still be needed, it will reduce the number of different roles involved. It therefore removes single points of failure in research reporting. Again, valuing this benefit is difficult as there is no direct cost saving. The potential cost to an institution of missing reporting deadlines could be explored.

Improve the institution's research profile

The ability to benchmark and undertake business intelligence should help institutions improve their success rate and hence their research profile – key for attracting funding, top researchers and postgraduate students. Harmonisation using the impact model and indeed benchmarking against other institutions' impact communications should help institutions to improve how they present their research outputs for maximum impact, again key to attracting and retaining key staff and students and marketing of an institution. It will additionally benefit PhD students, illustrating better the full cycle of research and its impact.

While it will be difficult to predict in advance the extent to which a research profile can be improved or the impact on staff and students, the value of retaining or recruiting new key staff and students could be estimated.

15.5.2 Dis-benefits of harmonisation for institutions

Ability to reuse data on research reporting inappropriately

Our workshops and sample interview with Research and Enterprise within a large institution did not identify any specific dis-benefits for large research-intensive institutions. Issues which were of concern to some research council representatives, such as inappropriate use of information or lack of authoritativeness, were of less concern to the institutional participants, largely as it was their data and by and large already in the public domain. That said, experience suggests that it would be prudent to implement cross-checks on data processes to prevent inappropriate usage.

Widen the gap between larger/research-intensive institutions and smaller/less research-intensive institutions

Larger/research-intensive institutions will have more data on which to base their business intelligence activities. This will give them a significant advantage, enabling them to out-compete smaller/less research-intensive institutions. This would negatively impact smaller/less research-intensive institutions; considerably widening the gap in income generation through research funding knowledge transfer and partnership activities.

Enablers and barriers to adoption of harmonisation by institutions

There are two key enablers for adoption of harmonisation by institutions – adoption by the research councils and incorporation of the harmonised research information models and dictionaries into commercial research information systems (CRIS). Without these, harmonisation is unlikely to be adopted.

Furthermore, as some funders pointed out, if harmonisation is primarily driven by the needs of the research council, funders' requirements might overshadow institutions' requirements. Significant additional cost could be incurred before an institution would realise any benefits, effectively inhibiting adoption.

The degree to which harmonisation is adopted within institutions is likely to vary. Smaller institutions or those without a strong research focus are unlikely to have the resources to update their systems to leverage the advantages that harmonised data offers. Their HR and Financial systems may still remain disconnected to their research information systems. This would make them less able to undertake business intelligence type activities or leverage all of the potential efficiencies.

15.5.3 Risks relating to harmonisation of research reporting for institutions

For large/research-intensive institutions

An analysis of the risks for large/research-intensive institutions if they implement harmonisation and if they do NOT as well as to the success of implementation is provided in table 15.4 below. As the analysis illustrates, the risk of not implementing harmonisation is likely to be much more significant than any risks associated with implementing harmonisation. Lack of a driver to implement if there is no push from the research councils or commercial systems available is also significant.

Risks for large/research intensive institutions in implementing harmonisation	Impact	Likelihood	Score	Mitigation
Inappropriate analysis of data leads to poor decisions or negative impact	2	2	4	Develop appropriate process safeguards and guidelines which make clear contextual limitations of data; use a champion to drive

				good practice
Harmonisation does not deliver the anticipated internal analysis benefits	2	2	4	Further develop basic harmonisation models and systems to enable desired internal analysis (this will incur additional costs)
Risks for large/research-intensive institutions in NOT implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Less able to compete for research council or funding agency money	4	5	20	Implement a harmonised approach
Less efficient research reporting	3	5	15	Implement a harmonised approach
Risks to successful implementation of harmonisation in large/research-intensive institutions	**Impact**	**Likelihood**	**Score**	**Mitigation**
Cost of implementation	3	2	6	Adopt a staged approach, implementing core requirements first
No driver for implementation	3	5	15	Work with funders to encourage them to move towards harmonisation; communicate benefits of harmonisation
No harmonised systems readily available	3	4	12	Encourage vendors to develop affordable harmonised CRIS solutions

Table 15.4: Analysis of the risks for large/research-intensive institutions associated with harmonisation

For small/less research-intensive institutions

An analysis of the risks for small/less research-intensive institutions if they implement harmonisation and if they do NOT as well as to the success of implementation is provided in table 15.5 below. As the analysis illustrates, the risk of not implementing harmonisation is likely to be much more significant than any risks associated with implementing harmonisation. Lack of a driver to implement if there is no push from the research councils or commercial systems available is also significant.

Risks for small/less research-intensive institutions in implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Inappropriate analysis of data leads to poor decisions or negative impact	2	2	4	Develop appropriate process safeguards and guidelines which make clear contextual limitations of data; use a champion to drive good practice
Harmonisation does not deliver the anticipate internal analysis benefits	2	2	4	Further develop basic harmonisation models and systems to enable desired internal analysis (this will incur additional costs)
Risks for small/less research-intensive institutions in NOT implementing harmonisation	**Impact**	**Likelihood**	**Score**	**Mitigation**
Less able to compete for research council or funding agency money	4	5	20	Implement a harmonised approach
Less efficient research reporting	3	5	15	Implement a harmonised approach
Risks to successful implementation of harmonisation in small/less research-intensive institutions	**Impact**	**Likelihood**	**Score**	**Mitigation**
Cost of implementation	2	4	8	Adopt a staged approach, implementing core requirements first
No driver for implementation	3	5	15	Work with funders to encourage them to move towards harmonisation; communicate benefits of harmonisation
No harmonised systems readily available	3	4	12	Encourage vendors to develop affordable harmonised CRIS

				solutions

Table 15.5: Analysis of the risks for small/less research-intensive institutions associated with harmonisation

15.5.4 Costs for institutions of implementing harmonisation

The costs for implementing harmonisation for institutions include: software; implementation of new data models or mapping; training; maintenance; and staff time. These costs will be more significant for smaller/less research-intensive institutions that are unlikely to have sophisticated CRIS or in-house systems already. Additional costs to realise wider indirect benefits are: new hardware; refactoring of legacy data and re-engineering processes; and creation and maintenance of registries.

15.6 Conclusions

15.6.1 Key findings

Benefits

For all the stakeholders the benefits of harmonisation are likely to outweigh the dis-benefits. All the stakeholders who would use harmonisation (as opposed to developing systems) are likely to see direct benefits of efficiency and improved quality and coverage of data. This will enable them to benchmark their performance and explore how to improve research outcomes. Of particular interest is that the harmonisation model could potentially help improve impact reporting, especially where collaboration is involved – something which has proved difficult to do effectively to date. Although harmonisation potentially affords process efficiencies and new analysis capabilities, to make the most of these organisations will need to further develop their information systems and legacy data.

While considerable potential benefits have been identified, valuing these is more problematic. The work undertaken as part of the business case for implementing CERIF wrapper provided some tentative details of efficiency savings; however, little work appears to have been carried out in institutions to estimate the value of benchmarking or business intelligence. Economic shadow pricing techniques could be used to find indicative pricing when developing the full business case.

Table 15.6 below provides illustrative examples of how the direct benefits associated with harmonising research reporting might be quantified and priced.

Direct Benefit	Quantification	Pricing
Improved quality of research reporting data	Staff grade and time saved in checking data	Hourly salary costs based on standardised FTE rates for relevant grade
Improved coverage of research reporting data	Number of additional fields populated	Based on shadow pricing of the value of the additional information
Improved business intelligence	Number and types of queries	Based on shadow pricing

data	undertaken; volume of reporting	approach currently being explored for Jisc business intelligence
Rationalised systems	System savings including: support, maintenance and running costs	Hourly salary costs for support based on standardise FTE rates for relevant grade Cost of maintenance contracts Running costs
More efficient data reporting	Staff grade and time saved in undertaking analysis tasks	Hourly salary costs based on standardised FTE rates for relevant grade
More efficient data analysis	Staff grade and time saved in undertaking analysis tasks	Hourly salary costs based on standardised FTE rates for relevant grade
More efficient business intelligence workflow	Staff grade and time saved in undertaking analysis tasks	Hourly salary costs based on standardised FTE rates for relevant grade

Table 15.6: Suggested measures and pricing approaches for direct benefits of harmonised research reporting

Dis-benefits

Some potential dis-benefits were identified. The potential to misuse and loss of granularity could be largely addressed by good information policies and participation in development of harmonisation specifications.

A more serious dis-benefit was the potential that harmonisation might widen the gap between research intensive and less intensive institutions. Similarly, charities could be disadvantaged if costs incurred prevented them adopting harmonisation. Indeed, the cost of implementation was a dis-benefit identified across the stakeholder groups.

Costings

The types of costs involved in implementing harmonisation are reasonably consistent across RCUK funders, charities and institutions. For small institutions and charities these are however a significant overhead and could limit uptake. While the business case for implementing CERIF wrapper provided some tentative costing for implementing the wrapper, interviewees and workshop participants all felt that the costs were unrealistically low. More work is required to develop a more realistic estimate of the cost of implementing harmonisation. Ideally this should be undertaken in a partnership with institutions and funders. Without their cooperation, it will be difficult to establish realistic costings.

Table 15.7 below provides illustrative examples of how the costs associated with harmonising research reporting might be quantified and priced.

Costs of basic harmonisation	Quantification	Unit pricing
Software	Staff grade and hours spend specifying system Number of systems commissioned or Hours spent in in-house development of system	Hourly salary costs based on standardised FTE rates for relevant grade Cost of commissioned systems or Hourly salary costs based on standardised FTE rates for relevant grade
Implementation of new data models or mapping	Staff grade and hours spent specifying and implementing new data models or mapping	Hourly salary costs based on standardised FTE rates for relevant grade
Training	Staff grade and hours spent by trainers Staff grade and hours spent by staff being trained	Hourly salary costs based on standardised trainer rates for relevant grade Hourly salary costs based on standardised FTE rates for relevant grade
Maintenance	Number and coverage of maintenance contracts	Maintenance contract costs
Staff time	Staff grade and hours spent on implementation not included above	Hourly salary costs based on standardised trainer rates for relevant grade
Costs of additional functionality such as business intelligence	**Quantification**	**Unit pricing**
New hardware and maintenance		Maintenance contract costs
Refactoring of legacy data	Staff grade and hours spent on refactoring	Hourly salary costs based on standardised trainer rates for relevant grade

Re-engineering processes	Staff grade and hours spent on re-engineering	Hourly salary costs based on standardised trainer rates for relevant grade
Creation and maintenance of registries	Staff grade and hours spent on creation and maintenance of registries	Hourly salary costs based on standardised trainer rates for relevant grade
Training	Staff grade and hours spent by trainers Staff grade and hours spent by staff being trained	Hourly salary costs based on standardised trainer rates for relevant grade Hourly salary costs based on standardised FTE rates for relevant grade
Staff time	Staff grade and hours spent on implementation not included above	Hourly salary costs based on standardised trainer rates for relevant grade

Table 15.7: Suggested measures and pricing approaches for costs of harmonised research reporting

Risks

In general, participants felt that the risks of not implementing harmonisation outweigh any risks associated with implementing harmonisation. Risks do not vary by size of funder; rather the likelihood of the risk becoming an issue does. It is different for institutions, with there being a risk of a widening research gap between large/research-intensive institutions and small/less research-intensive ones.

There is also a significant risk that a partially harmonised model developed by the research councils is adopted, resulting potentially in a one supplier solution, which does not meet the reporting and business intelligence needs of the wider HE research community. While such a solution might lead to a high level of uptake as it is required for reporting on research council-funded research, without appropriate governance, other funders could find it difficult to adopt or have no voice in submitting changes that would make it suitable for their requirements.

Finally, there is also a risk that lack of sector-wide leadership on harmonisation may mean that a non-optimal approach to harmonisation emerges as well as uptake being limited due to lack of support and benefits realisation management.

Barriers and enablers

Cost of implementation and lack of a clear driver for implementation were viewed by far as the biggest barriers to uptake of implementation. Harmonisation of research reporting across all the research councils was viewed as the biggest enabler followed by development of CRIS systems that support the harmonised models and dictionaries. Adoption by the research councils of their internal system rather than a UK-wide harmonisation model that supports the requirements of all the stakeholders in research reporting could potentially act as a barrier to the full realisation of benefits across UK HE. A sector-wide leadership role which seeks to drive an optimal approach to harmonisation (cost-effective while meeting the whole sector needs) as well as provision of support information and benefits realisation management would significantly aid both effectiveness and uptake.

Recommendations and next steps

Research reporting is a key part of the research ecosystem as it helps manage the return on investment in research as well as disseminating research outputs. Many stakeholders are involved including: the government as the funder of research councils, as well as the councils themselves and other funders; research organisations and individuals; others such as vendors of research reporting systems; and business and society as the ultimate benefactors of the research. **The business case for harmonisation of research reporting, therefore, needs to be situated within this complex innovation ecosystem, illustrating the business case for each of the key stakeholders as well as how it will contribute to increasing economic and social benefits to business and society.**

The business case needs to consider three options:

The research councils are currently investigating ways to increase harmonisation in the reporting of research outcomes. Once their way forward has been announced, a more accurate business case for adoption of a harmonised approach to research reporting across the UK HE sector could be explored.

There is a wide range of stakeholders involved, many of whom would gain direct benefits from the harmonisation of research reporting through a core information profile of common fields together with supporting CERIF mappings and dictionaries of terms.

This case would need to not only compare the value the benefits against the costs of implementation for the key stakeholders involved, but it would also need to address the dis-benefits as well as identifying the enablers that need to be put in place if harmonisation is to be adopted and benefit the whole of UK HE. Development of such a business case would require the cooperation of key stakeholders – in particular representative research councils, research intensive and non-research intensive institutions, charities and CRIS suppliers. Without reliable data from these stakeholders it is highly unlikely that a sufficiently detailed business case could be developed.