Using CRIS to power research data discovery

Alex Ball[1] Christopher Brown[2] Laura Molloy[3]
Veerle Van den Eynden[4] David Wilson[3]

12 May 2015

Abstract

In the UK, Jisc is developing a national Research Data Discovery Service (RDDS) in partnership with the Digital Curation Centre and the UK Data Archive. In the first phase of the project (2013–2014), the team set up an experimental service based on the ORCA software developed by ANDS for Research Data Australia. Unlike the Australian version, the UK instance harvested metadata in a variety of formats and performed crosswalks to convert it to the required format. Since the repositories and the metadata standards they used were focused on the datasets themselves, some useful information about researchers, projects and funders was missing.

In the second phase of the project (2014–2016), the team is looking at ways of filling those gaps by harvesting not only from data repositories but also from CRIS instances. We present a metadata crosswalk from CERIF to RIF-CS – the standard used by ORCA – and highlight the issues and challenges raised by using it in the context of the RDDS. We also consider the positive effect that a more complete network of information has on the discoverability of research data.

Contents

[1]DCC/UKOLN, University of Bath, [2]Jisc, [3]DCC/HATII, University of Glasgow, [4]UK Data Archive.

1 Vision

The first thing to stress is that we are focusing on *UK research data*. So far we've found it convenient to interpret that as data that can be found in discipline-specific data centres and institutional data repositories in the UK. We may well look further afield in future, but it is a good starting point.

§ The second thing is that this is a *discovery service*, not a super-repository. We will not host any data ourselves, but will make data as visible as possible wherever it happens to be. . . . important for long-tail research, and for cross-disciplinary work; relevant data might be scattered across 170 UK HEIs and a dozen data centres, and no-one wants to search through them all individually.

§ Ultimate goal is to encourage *data sharing and reuse*, so that the data has maximum *impact*, researchers get due *credit* for it, funders get better *value for money* and we have a more complete and higher quality scientific *record*.

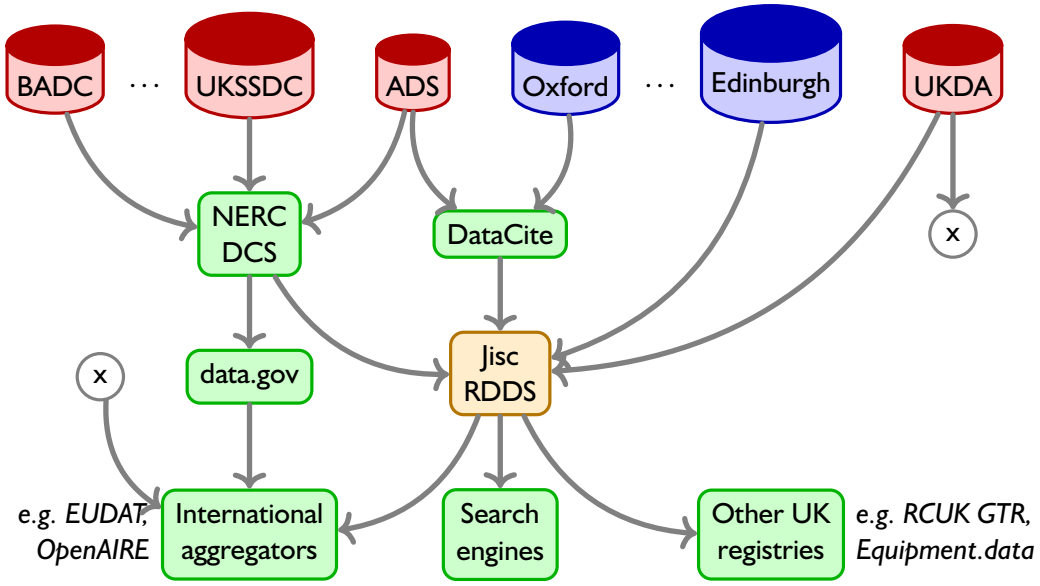


Figure 1: Place of the RDDS in the repository landscape. Note that many repositories will also contribute directly to international aggregators.

¶ Of course, there are already services doing similar things – OpenAIRE and EUDAT spring to mind – but we have the potential to complement rather than compete with them (see Figure 1), by:

- collating records from both data centres and institutional repositories;
- normalising and deduplicating, to provide a unified search interface;
- ultimately make the records visible in other places researchers might look.

2 Phase 1

In Phase 1 of the project, we built a pilot service to give us a better idea of the technical challenges and stakeholder requirements.

- Discovery service based on ORCA (Research Data Australia)
  - *we knew it worked well in Australia;*
  - *we at the DCC were already working with the developers to make the code more portable;*
  - *we liked the way it was search engine friendly, and provided citation and rights information up front.*
- Participating data repositories:
  - 9 universities: Edinburgh, Glasgow, Hull, Lincoln, Leeds, Oxford, Oxford Brookes, St Andrews, Southampton
    All of these had or were developing data repositories, and had data records we could harvest.
  - UKDA
  - Archaeology Data Service
  - 7 NERC Data Centres: BADC, BODC, EIDC, NEORC, NGDC, PDC, UKSSDC
    These gave us a wide disciplinary range, plus we could cheat a bit because all except the UKDA contribute to the NERC Data Catalogue Service.
- Harvested metadata in six standard formats
- Harvested metadata according to three protocols:
  - native XML export
  - OAI-PMH
  - CSW (Catalogue Services for the Web)
  - (but we had to cheat a little bit with the last two)

Finally, we set up accounts for all the participating repositories and asked them to try importing records into the registry. Which they did.

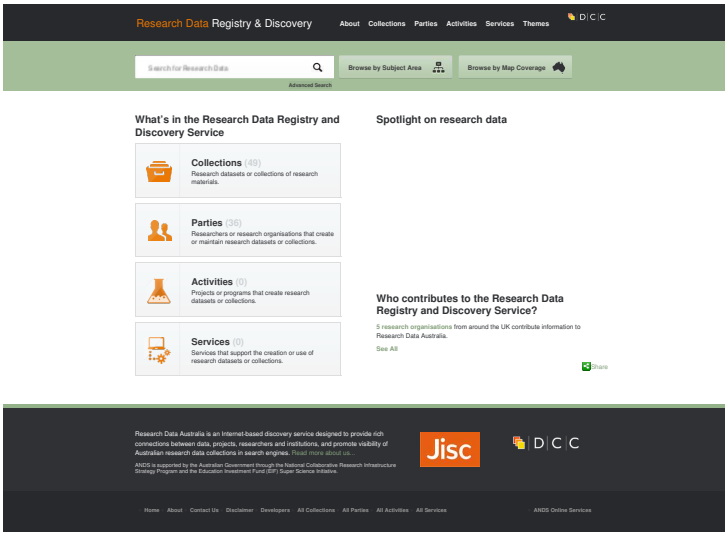


Figure 2: The pilot UK Research Data Registry and Repository Service, URL: `http://rdrds.cloudapp.net/`

¶ So here is what the pilot service looks like (see Figure 2). There are still a few remnants of Research Data Australia . . . more work to do on making the code portable. . . . If you

do visit it there isn't much to see because most of the records that have been imported have not been made public. I think in some cases the contributors tried using the OAI-PMH harvester instead of manually feeding in an OAI-PMH response. We had to feed the responses in manually because at that time it could only harvest records in…

3 RIF-CS

…RIF-CS format, which is used internally by the service.

- Profile of ISO 2146 (Information and Documentation – Registry Services for Libraries and Related Organizations)
- Optimized for collection services registries
- Maintained by ANDS: see `http://services.ands.org.au/documentation/rifcs/1.6/guidelines/rif-cs.html`
- 'Gateway drug' for CERIF?

It is not as detailed as CERIF, but it shares some of same philosophy: it moves you away from thinking in terms of a single flat metadata file and starts you thinking about relationships between different entities.

¶ There are only four entities, but they are specialised with types (Figure 3).

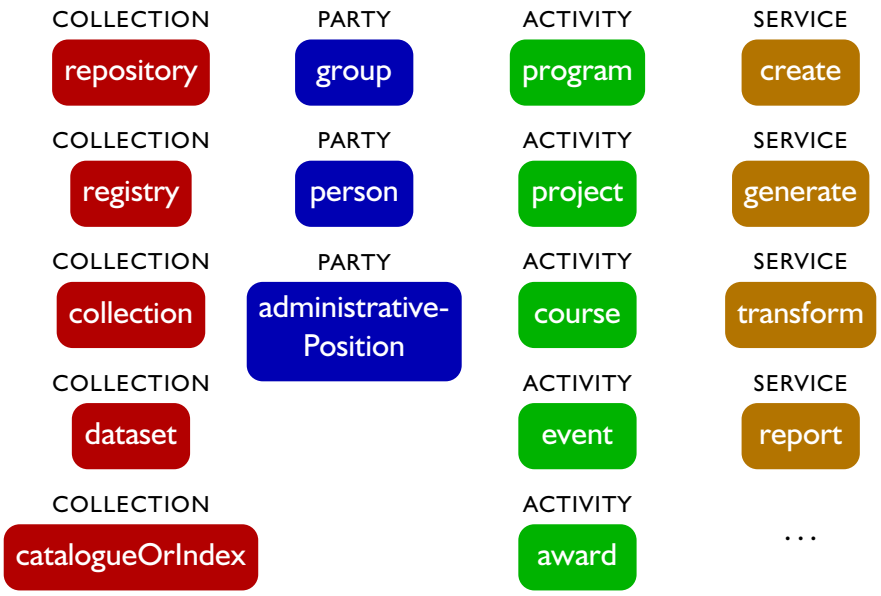


Figure 3: Example entities from the RIF-CS data model

With these you can build up a quite detailed network of records (¶ Figure 4)….Not just about elegance or efficiency: these relations are also browsing pathways.

This is all very well, but RIF-CS is not very well known or supported. As this was just a pilot, we couldn't expect our partners to implement it and send us ready made records. It was down to us to perform the crosswalks centrally.

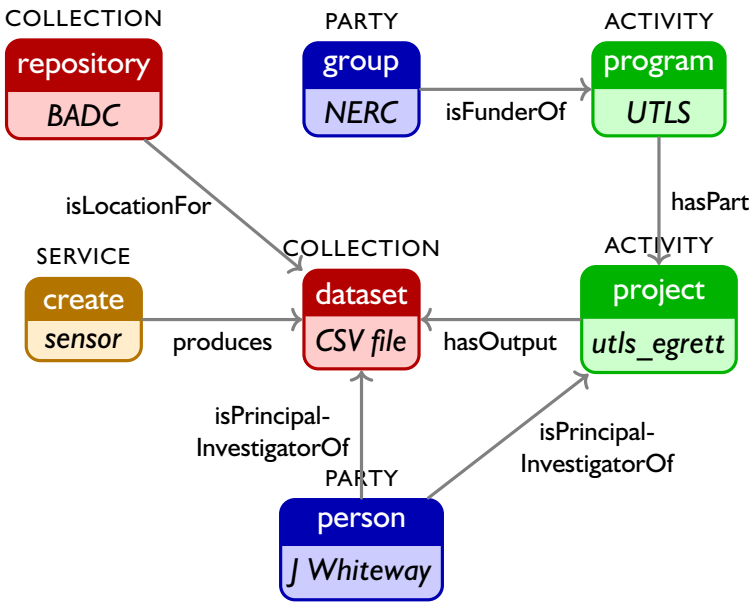


Figure 4: Example set of related objects

¶ So we had to write crosswalks to harvest records in formats that were supported. OAI DC is a fallback that most repositories should support, but we also wanted to benefit from more detailed metadata that many repositories might be able to provide (Figure 5).

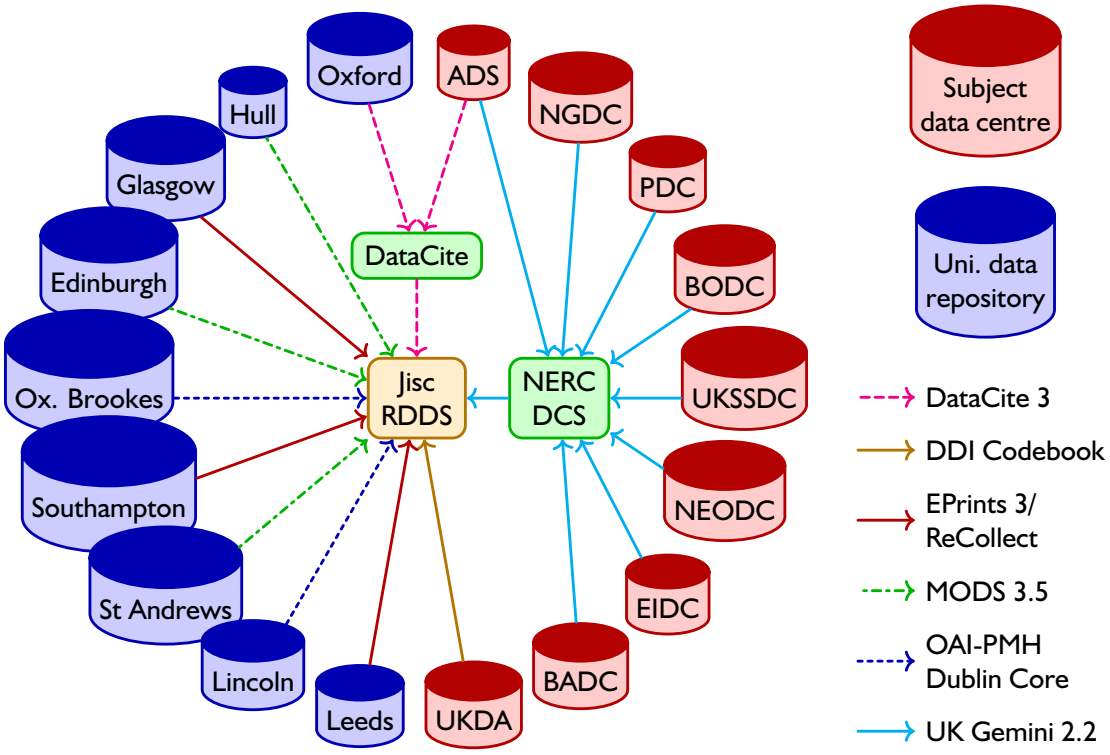


Figure 5: Metadata flows into the Phase 1 pilot RDDS.

For the most part, this worked well. We could generate fairly detailed dataset records, even from OAI DC, but the problems came when we tried to generate records for the parties involved: the data creators and the funders (¶ Figure 6).

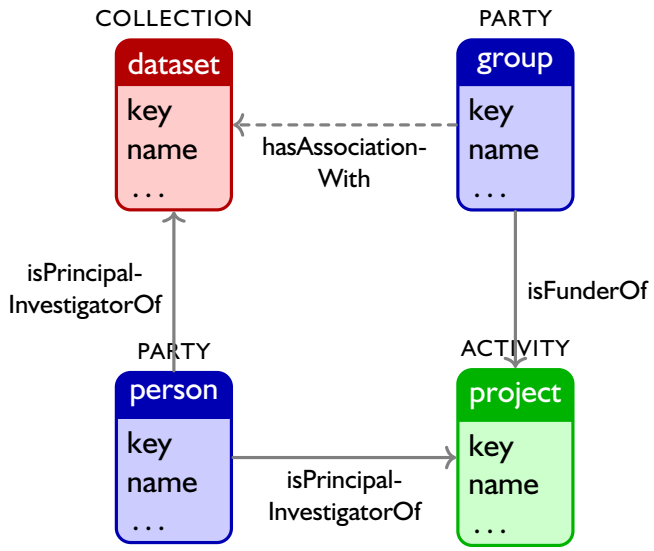


Figure 6: Most dataset records omit the project information, making it hard to connect funders with datasets.

The information about creators was often sparse; very few contributors could give us identifiers for them, and there simply wasn't enough information to be able to generate identifiers reliably. By that I mean one ID per person and one person per ID. The other issue was that, while some contributors could tell us who had funded a dataset, we couldn't express that relationship directly in RIF-CS, at least not without resorting to the rather vague 'hasAssociationWith' relation.

§ To fix this, we will need more information about people and projects. This is where the ability to harvest information from the CRIS would be very useful indeed.

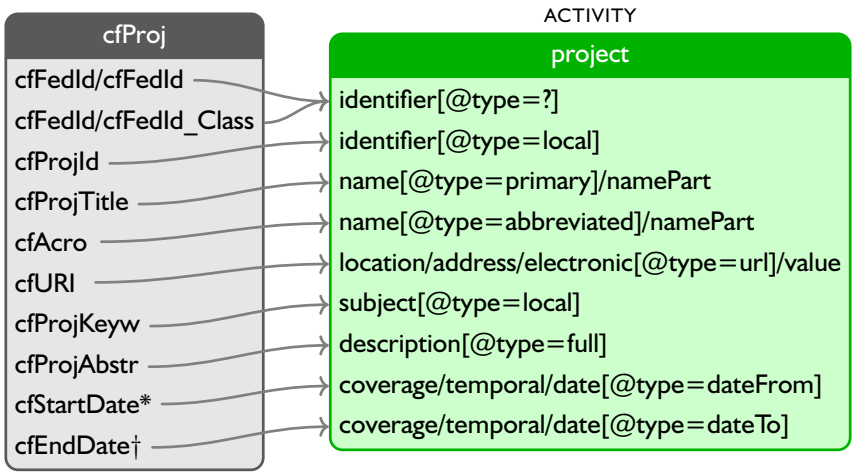
To this end, I would like to run by you this preliminary mapping from CERIF 1.6 – particularly the profile of it developed for OpenAIRE[1] – to RIF-CS.

4 Mapping from CERIF to RIF-CS

I'll start with the mapping for projects, which were absent from our Phase 1 pilot (Figure 7). I should explain that as well as the properties you see up here, all RIF-CS registry objects, or records, can specify *related objects* and *related information*. Related objects are other records in the database, while related information does not have a matching record. Its easier if I show those relationships later.

For now, you can see that we can handle multiple identifiers, global and local ones. I haven't shown the one the registry software would generate itself. We can also take the long and short form of the project name, its home page, a description of it, keywords that would hopefully identify its disciplinary scope, and the time period for which it was (or is) active. The latter appears to have become quite complicated to find, so I guess we'll need to cope with many eventualities.

[1] OpenAIRE Guidelines for CRIS: `https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_CRIS`



* deprecated; see also 'Awarded' `cfProj_Class/cfStartDate`, 'Funder' `cfProj_OrgUnit/cfStartDate`.

† deprecated; see also 'Awarded' `cfProj_Class/cfEndDate`, 'Closed' `cfProj_Class/cfStartDate`, 'Funder' `cfProj_OrgUnit/cfEndDate`.

Figure 7: From CERIF to RIF-CS: Projects

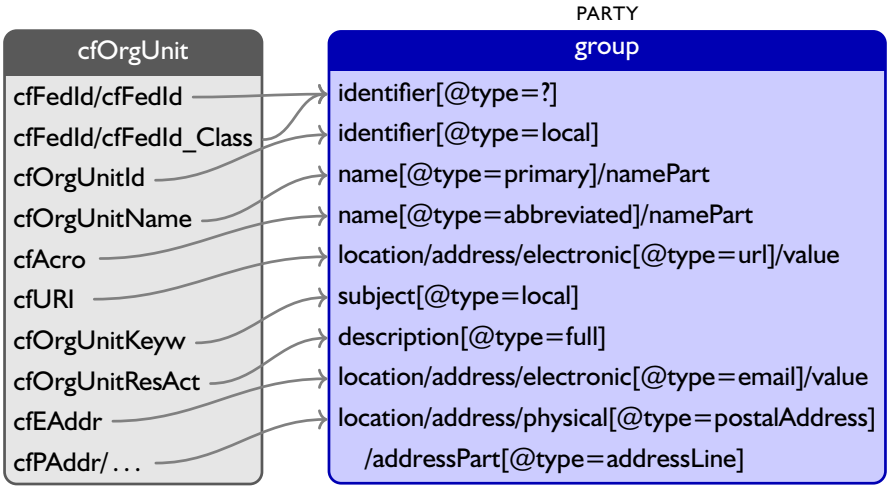


Figure 8: From CERIF to RIF-CS: Organisations

¶ Next is the mapping for organisations, which we'll need for funding bodies and researcher affiliations (Figure 8). As you can see, it's a very similar mapping, the main difference being instead of dates at the bottom we've mapped contact details. I've also assumed on the basis of the sample files I've seen that Organisation Unit Research Activity will give a narrative description of the organisation. To be honest the most important part is the ID, as we are most likely going to write these records ourselves and will be looking to recognise these records rather than harvest them.

We are much more likely to harvest records relating to people (¶ Figure 9). The mapping falls into much the same pattern, and again we are particularly interested in IDs. But the main point of interest here lies in names. CERIF is of course set up to deal with a multiplicity of names through `cfPersName_Pers` relations, and can even timestamp their period of applicability. RIF-CS isn't quite that expressive but it does give us primary and alternative names.

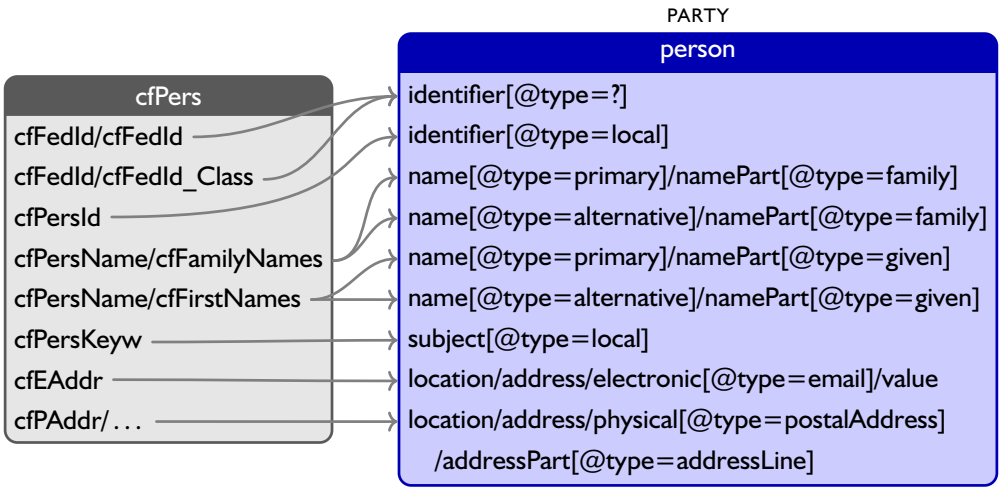


Figure 9: From CERIF to RIF-CS: Data creators

When we do the mapping for real, we may have to choose the primary name from several `cfPersName_Pers` relations. If one is classified as preferred, we'll choose that, otherwise we'll pick the most recent.

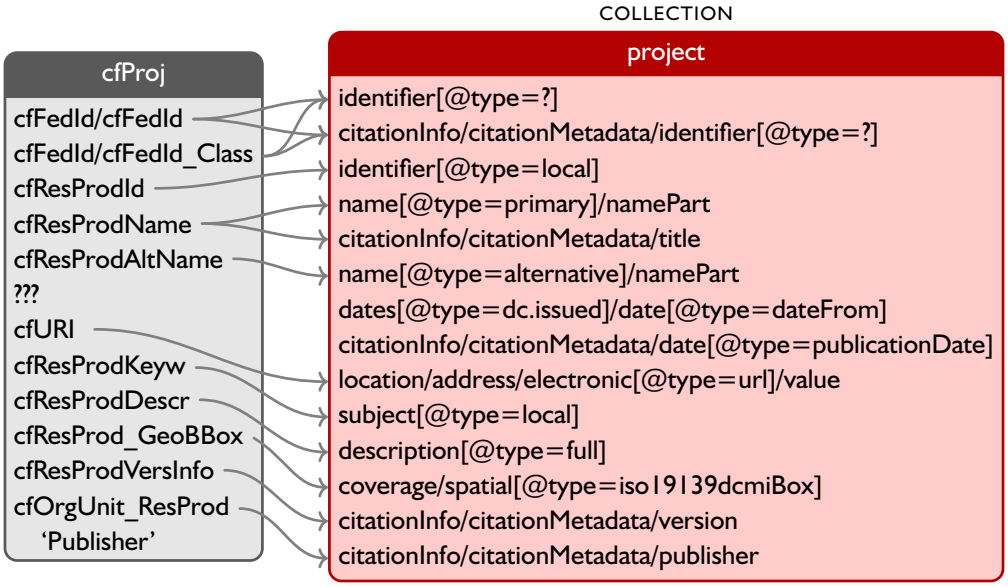
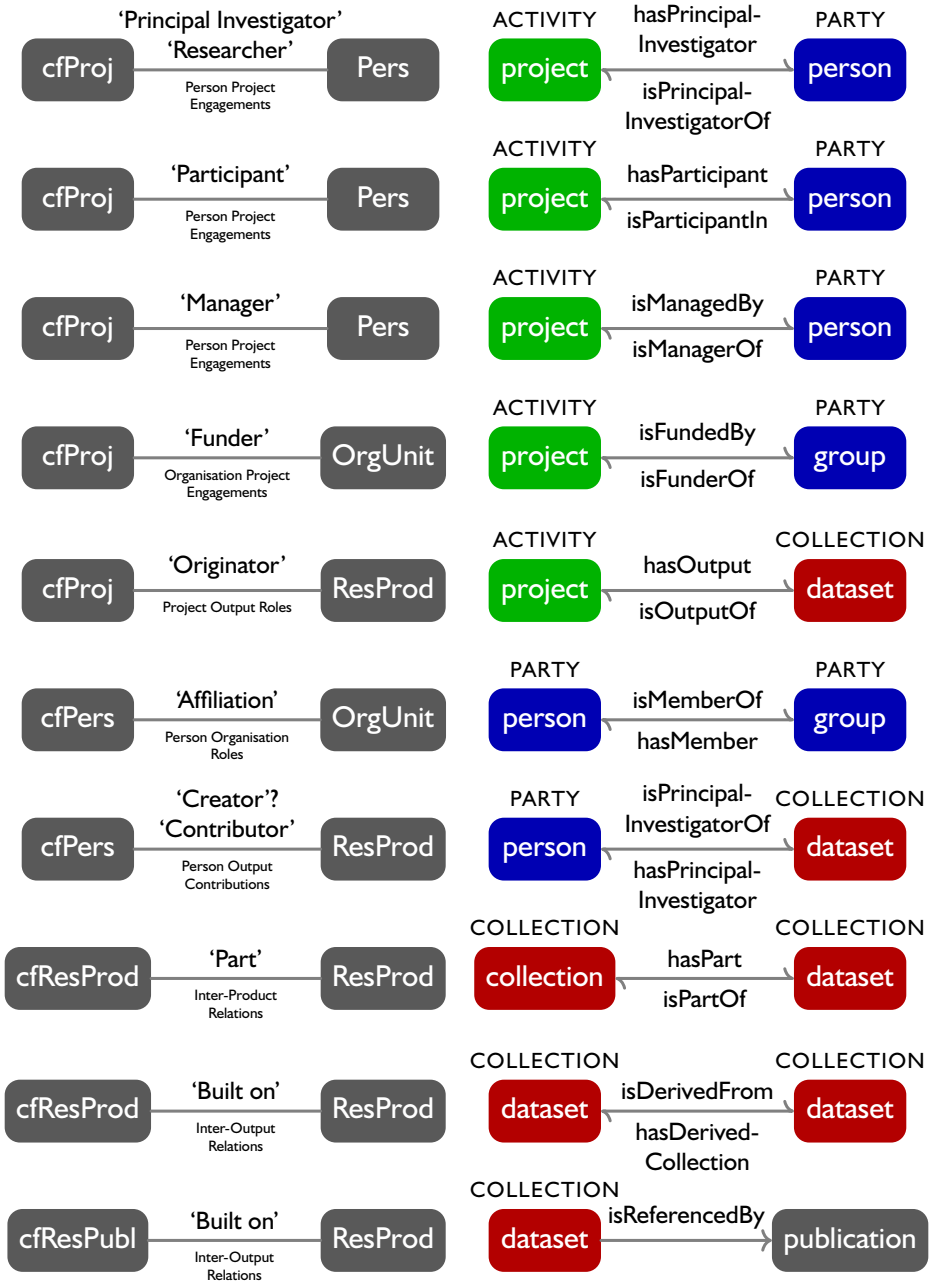


Figure 10: From CERIF to RIF-CS: Datasets

So far we have been harvesting from dedicated data repositories, but we anticipate that institutions will increasingly want to manage their research data from their CRIS. Therefore we've also put together a mapping for datasets (¶ Figure 10). It looks much busier because some information is duplicated, albeit in a slightly different form, in a citation information section used for building a sample citation. One of the required components of that section is a date of publication or release. It is not immediately clear to me where to find that information. I didn't have room on the slide, but we would also be interested in including temporal coverage (e.g. collection period of observation), and any rights information (licences, access restrictions, usage restrictions). We would also value, and this goes for all the entities, knowing which controlled vocabulary, if any, has been used for the keywords.

¶ Lastly, I promised you a mapping of relationships, and so here they are. I'm quite pleased about the correspondence we're able to get between the two schemes. One thing to note is that if you look at the term definitions in RIF-CS, the P.I. relations are actually used for any researcher: I suspect this is a case of a term mutating away from its original definition.



RIF-CS could also express

- the party that manages the dataset;
- the party that owns the dataset;
- a publication that cites the dataset;
- a publication that documents the dataset;
- a publication to which the dataset is a supplement.

That's the mapping. I offer it as a conversation starter, really, as it is not confirmed at this point whether we will go ahead with RIF-CS or use something else. But to finish with, I'd just like to reflect on what it will mean to us if we can get it working.

5 Phase 2, with added CRIS?

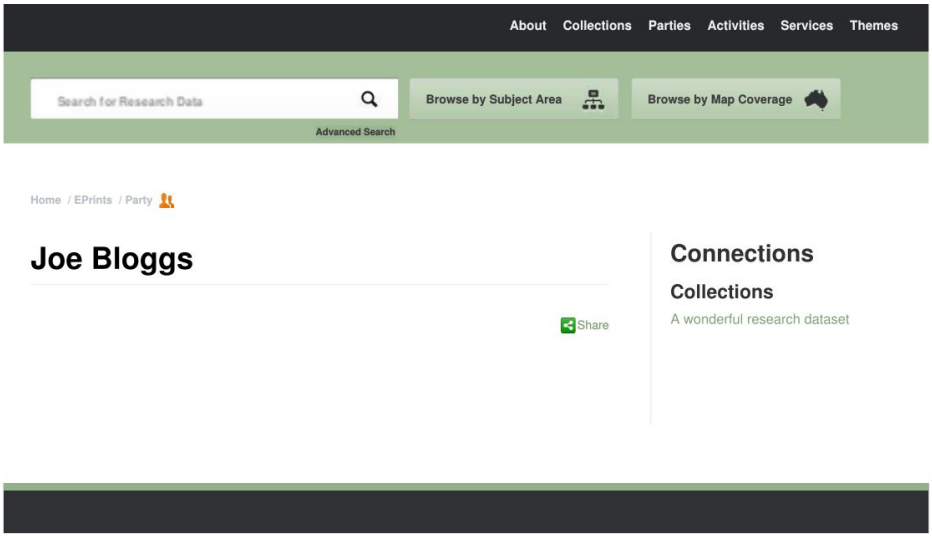


Figure 11: Example party record: without CRIS

Figure 11 shows the kind of record we generated in Phase 1 for data creators. We have just a name and a related dataset. Disappointing. At the very least we'd want something a bit richer, ¶ as in Figure 12.

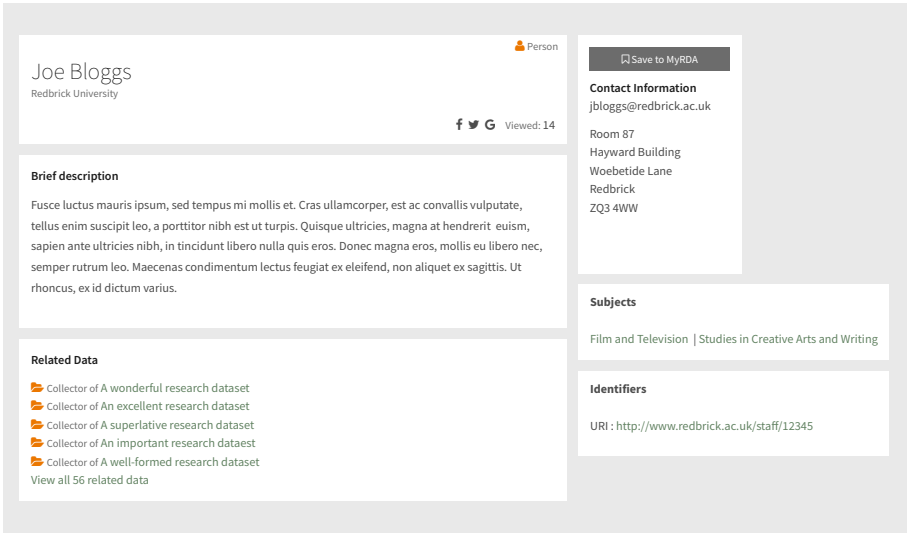


Figure 12: Example party record: with CRIS?

We want people to be able to click around on here on the projects this person has worked on, to see what other data came out of the same activity. We want to use the web of connections to report on what datasets have come out of a particular university, or were made possible by a given funder. In the long term, we'd want researchers to be

able to track where their data has been mashed up into other datasets, and how many papers have come out of the data they have shared.

We can only do this by collecting a complete network of information, and while we might not be able to everything we need from CRIS at the moment, what we can get will be a great help in meeting our goals.

Alex Ball[1] , *Christopher Brown*[2] , *Laura Molloy*[3] , *Veerle Van den Eynden*[4] , *David Wilson*[3].
[1]*DCC/UKOLN, University of Bath* , [2]*Jisc* , [3]*DCC/HATII, University of Glasgow* , [4]*UK Data Archive*.

The DCC is supported by Jisc.

For more information, please visit `http://www.dcc.ac.uk/`

UK RDDS Phase 2 home page: `http://www.jisc.ac.uk/rd/projects/uk-research-data-discovery`