RDA-WDS Publishing Data Interest Group

Cost Recovery for Data Centres Working Group Case Statement

Working Group Charter

Objectives

Basic funding of data infrastructure may not keep pace with increasing costs[1]. Therefore, there is a need to consider alternative cost recovery options and a diversification of revenue streams. In short: who will pay for public access to research data?[2]

This Working Group proposes to make a contribution to strategic thinking on cost recovery by conducting research to understand current and possible cost recovery strategies for data centres. The Working Group will pay particular attention to data centres' involvement in data publishing activities and examine such initiatives as a potential source of alternative revenue.

The Working Group will produce a report providing conclusions and recommendations about the potential appropriateness of different cost recovery models to different situations and the potential of data publication initiatives fitting into a cost recovery strategy. The Working Group will also

[1] Sustaining Domain Repositories for Digital Data: A White Paper http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

[2] F. Berman, V. Cerf, 'Who will pay for public access to research data?' in Science Magazine, vol.341, p.616-617

contribute its findings to the combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups.

These deliverables will build on five areas of work:

1. A summary of current work on cost models[3];
2. A survey of funding policies specifically relating to how the costs of data availability/publication may be recovered;
3. A survey, by means of a questionnaire and case studies, of various existing approaches to cost recovery/business models;
4. A survey of other stakeholders (publishers, researchers) to understand their position and policy in relation to charging models and their role in the publishing process;
5. The outcomes of the Working Group on Workflows.

Value proposition: Cost recovery for Data Centres

The challenge: cost recovery and sustainability of public data products.

A number of initiatives examining data citation and data publication aim to integrate data more effectively with the process of scholarly communication and the ‘record of science’. A vision of the future in which ‘data papers’ have scholarly currency, in which data can be accessed and visualized directly from the online literature requires partnerships between publishing platforms and data centres (e.g. [4]). This vision also demands that data centres providing access to published datasets should have sustainable business models.

A lot of work is going on to understand the costs of maintaining long-term accessibility to digital resources, to identify different cost components and based on this to develop cost models[5]. However, in a broader context—which considers data as part of research communication—the identification of costs and development of cost models addresses only part of the problem. In times of tightening budgets, it is important to address the challenge of ensuring the sustainability of data centres - and considering this in the context of the broader processes for data publication. Many established national and international data centres have reliable sources of income from research funders. However, these sources of income are generally inelastic and may be vulnerable. There is concern that basic funding of data infrastructure may not keep pace with increasing costs. Therefore, there is a need to consider alternative cost recovery options and a diversification of revenue streams.

This Working Group proposes to make a significant—but achievable—contribution to strategic thinking in this area by conducting research to understand current and possible cost recovery strategies for data centres. We will pay particular—but not exclusive—attention to data centres’

[3] In this document, we will differentiate between two aspects: ‘cost models’, i.e., a description of what different aspects of data curation and storage cost, and ‘cost recovery’, i.e., models and ways in which data centres can charge for their services. We consider both components to be within scope of this working group, but the focus of the questionnaire and the testing is to provide a practical overview and advice on various cost recovery models.

[4] Science as an open enterprise, The Royal Society Science Policy Centre report 02/12, Issued: June 2012 DES24782, ISBN: 978-0-85403-962-3, The Royal Society, 2012

[5] www.life.ac.uk, www.costmodelfordigitalpreservation.dk, www.beagrie.com/jisc, http://brtf.sdsc.edu/, http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2, http://www.alliancepermanentaccess.org/index.php/aparsen/, http://4cproject.net/

involvement in data publishing activities and examine such initiatives as a potential source of alternative revenue.

By means of a questionnaire and a set of case studies, this working group will shed light on data centres' current practice of cost recovery and identify possible opportunities for data centres looking to diversify income streams. A number of important questions will be considered, including:

- What cost recovery models are currently being employed by data centres?
- What trends are perceived by data centres with regard to the vulnerability of funding and what are the possible responses to diversify income streams?
- What cost recovery models are available within current, largely grant based funding of research?
- What cost recovery strategies are available while maintaining a commitment to open access to research data?

The principal activity of the WG will be to survey a set of data centres and provide a group of case studies addressing these questions in detail, for use within the test environment developed by other RDA Working Groups. This work will build upon existing work on cost models and funder's policies, as they relate to cost recovery of data curation costs from research projects. These aspects of the work will help the WG analyses how the costs of particular activities may be covered, and to what extent it is possible to hypothecate charges and encourage clarity around who pays for what. However, the principal focus will be on understanding the alternative options available for cost recovery and diversification of revenue streams for data centres. There are various options available and the involvement of data centres with 'data publication' initiatives is a significant new development that will be considered in this study.

Summary of the cost components of a data publication:

- Ensuring a publishable data product—annotation, metadata, codebooks etc.—is argely the responsibility of the researcher.
- Quality assurance and review process is conducted in some instances by the publisher but also, largely, by the data repository.
- Long-term preservation—archive and services for access— is largely the responsibility of the data repository.

Scope and Terminology

The working group has a realistic objective of conducting a survey among a limited set of data centres in addition to a small number of carefully selected case studies. In order to be as informative as possible, this research will be narrow and deep. Additional funding will be sought to allow the expansion of this activity, the involvement of a greater number of data centres in the survey and as case studies.

Deliverables

The working group will produce a report providing conclusions and recommendations about the potential appropriateness of different cost recovery models to different situations and the potential of data publication initiatives fitting into a cost recovery strategy.

With respect to the implementation of the recommendations from the final report, we choose the following approach. The recommendations will produce a number of scenarios for cost recovery. These scenarios will contain criteria to determine their appropriateness to different situations. For each scenario a use case will be tested in practice by one of the stakeholders involved (digital repository/publisher/research institution/etc.). The cost recovery model will be embedded in their services and their administrative and financial workflows.

Next to this, the Working Group will contribute its findings to the combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups. We will match their workflow reference models with our cost recovery scenarios: what are the practical implications of the different scenarios on the workflows?

These deliverables will build on five areas of work:

1. A summary of current work on cost models—the WG will simply provide a summary of existing work (4C, ICPSR, APARSEN, etc.)—will inform our own analysis.
2. A survey of funding policies specifically relating to how the costs of data availability/publication may be recovered—the WG will provide a brief summary of funder expectations or rules governing the funding of data deposit from research grants. This will build on some existing work conducted by the Knowledge Exchange, DCC and others. We will follow up with specific approaches to certain funders to understand the principles underpinning the policies and any possible changes in direction.
3. A survey—by means of questionnaire and case studies—of various existing approaches to cost recovery/business models. This will be the core new activity of the Working Group. The WG will survey members of the World Data System, holders of the Data Seal of Approval, members of ICPSR and other established data centres.
4. On the basis that possible cost recovery options include pay-to-deposit and pay-to-access, other stakeholders—funders, publishers and researchers—will be surveyed to understand their position and policy in relation to charging models and their role in the publishing process.
5. The outcomes of the Working Group on Workflows. This Working Group will produce a classification of a representative range of workflow models. In each case, the varying stakeholders and their different roles and responsibilities will be identified as well as the likely associated resource and cost implications.

Who will benefit

The principal beneficiary of this work will be data centre managers who will have an insight into alternative options for cost recovery, substantiated by case studies. Other stakeholders in data publication will benefit from a clearer insight into the relationship between policy, funding and cost recovery. The ultimate contribution of the WG will be to present the community of stakeholders with examples of how sustainability of data infrastructure for publication may be achieved.

Engagement with existing work in the area

There is a significant existing body of work on cost models. The European 4C project has undertaken a significant task of synthesis in this area. Other initiatives on costs and cost models include Knowledge Exchange and APARSEN.[6] The Working Group will closely monitor the developments and outcomes of this work and it will feed into our summaries of cost models and the survey of policies.

We are not aware of other work in the specific area of cost recovery, hence the need for this Working Group. There are a number of practical initiatives for data publication exploring new business models for data repositories, for example, Dryad.[7] Similarly, a number of established data centres (DANS in the Netherlands, ADS in the UK) are exploring new or supplementary approaches to cost recovery and these will be included among our case studies.[8]

The work of other Working Groups within the RDA-WDS Data Publication IG will be taken into account, in particular the Data Publication Workflows WG. The outcomes of our Working Group will feed into the Workflows WG.

We will also pay attention to the RDA Data Foundations and Data Certification Groups and the work of these groups will inform the framework used. Nevertheless, the focus of this WG is very much on understanding existing and possible approaches to cost recovery and this is a niche that is not being explored by other groups.

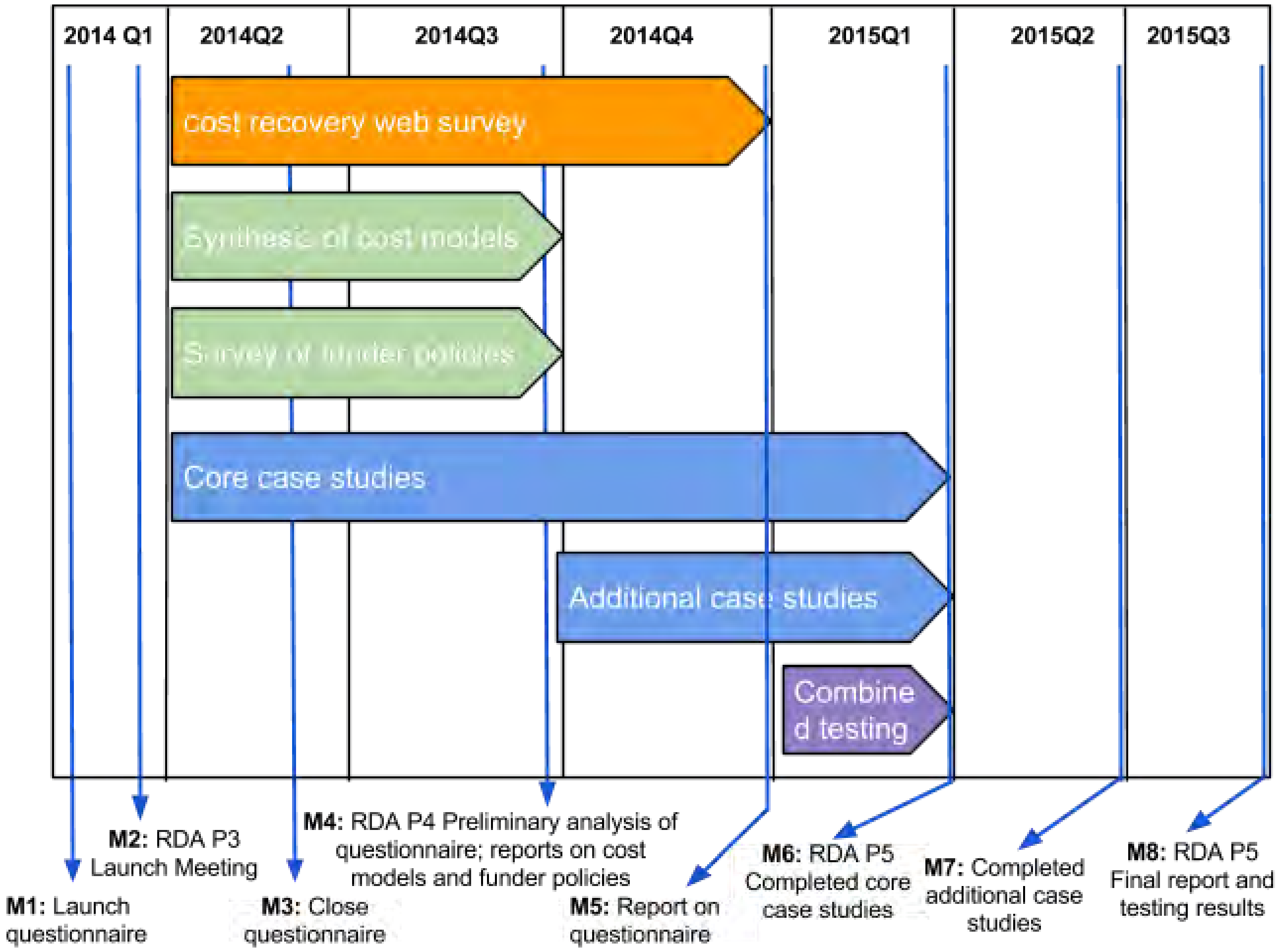
Work Plan

Deliverables and activities

1. Summary report of existing cost models to identify key cost points.
   - Summarize current cost models with specific consideration of how cost points identified might relate to cost recovery (or not).
2. Summary report of funder data policies insofar as they relate to cost recovery. This report will identify what funder policies say about the scope for cost recovery of data infrastructure from research grants (link with Knowledge Exchange work on funding).
   - Survey funder data policies insofar as they relate to cost recovery.
3. Questionnaire to survey cost recovery in a sample of data centres.
   - Design and set up questionnaire.
   - Compile list of target participants.
   - Conduct survey.
   - Analyse results.
4. Case studies from a selected set of data centres, providing more detailed analysis of cost recovery approaches.

[6] http://4cproject.net/, http://www.knowledge-exchange.info/, http://www.alliancepermanentaccess.org
[7] http://datadryad.org/
[8] www.dans.knaw.nl, http://archaeologydataservice.ac.uk/

    - Set up initial set of case studies (e.g. ADS, Dryad, DANS, CCDC)[9]
    - Results of questionnaire may be used to identify additional case studies. Case studies will provide detailed examples of cost recovery models; the rationale behind the choices made; the experience of the model and the current estimate of its appropriateness and likelihood of success.
5. Report on potential cost recovery models
    - On the basis of the questionnaire, case studies, and stakeholder engagement, the Working Group will prepare a report summarizing available cost recovery models and identifying the most likely ways in which providers of data infrastructure may diversify their income streams.
    - Which cost recovery strategies are most likely to be transferrable?
    - Are there patterns indicating that particular models may be more appropriate for certain institutions?
    - To what extent does ‘data publication’ offer an additional opportunity for cost recovery?
6. Combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups
    - match the workflow reference models with cost compensation models
    - identify the different cost components and the potential cost recovery options


2014 Q1
2014Q2
2014Q3
2014Q4
2015Q1
2015Q2
2015Q3
cost recovery web survey
Synthesis of cost models
Survey of funder policies
Core case studies
Additional case studies
Combine d testing
M1: Launch questionnaire
M2: RDA P3 Launch Meeting
M3: Close questionnaire
M4: RDA P4 Preliminary analysis of questionnaire; reports on cost models and funder policies
M5: Report on questionnaire
M6: RDA P5 Completed core case studies
M7: Completed additional case studies
M8: RDA P5 Final report and testing results

[9] http://www.ccdc.cam.ac.uk

Milestones

Before 18 January 2014: submission of RDA Case Statement.

- **M1:** 1 March 2014: launch questionnaire (continual reminders)
- **M2:** 26 March 2014, WG Launch Meeting at RDA Plenary in Dublin.

1 month - April 2014: start synthesis of cost models and survey of funder policies; initiate core case studies.

- **M3:** 30 June 2014: close questionnaire

6 months - end Sept 2014: preliminary analysis of questionnaire; reports on cost models and funder policies.

- **M4:** 22 September 2014 (RDA P4): present preliminary analysis of questionnaire; reports on cost models and funder policies

9 months - end December 2014: prepare report on questionnaire

- **M5:** 31 December 2014: Present report on questionnaire

12 months - end March 2015: (core) case studies;

- **M6:** 31 March 2015: present completed (core) case studies;

15 months - end June 2015: (additional) case studies and start combined testing within the four Data Publishing Working Groups

- **M7:** 30June 2015: Present completed (additional) case studies, start combined testing within the four Data Publishing Working Groups

18 months - end September 2015: prepare the final report and testing results

- **M8:** 30 September 2015 (RDA P5): Present final report and testing results

Project Management

The Working Group will have regular web meetings for working group members every six weeks.

Face-to-face meetings will be attached to the RDA Plenaries in Dublin in March 2014, in Amsterdam in September 2014, in March 2015 and in September 2015.

Other face-to-face meetings may be arranged if necessary, either for the whole group or for those working on specific work packages.

The Working Group leaders will participate in the regular meetings of the WDS/RDA Data Publishing Interest Group. These meeting will be used to keep the project on track, to monitor progress and resolve any differences of opinion.

Adoption Plan

Dissemination and Engagement

- Preliminary analysis of the questionnaire and the reports on cost models and funder policies will be disseminated at the RDA Plenary in Amsterdam in September 2014 and at SciDataCon in New Delhi in November 2014.
- The core case studies will be discussed at the RDA Plenary in Spring 2015.
- The final report will be delivered at the RDA Fall Plenary 2015.
- The publishers will be involved through the Data Working Group of STM, the ALPSP International Conference and the APE.

- The Working Group will also engage with generic meetings for data infrastructure providers, such as the WDS meetings, DataCite, etc.
- We will also target subject-focused conferences relating to the case studies: these are likely to cover social sciences and humanities (DANS), archaeology (ADS), crystallography (CCDC), life sciences/ecology (Dryad), European Geophysical Union/American Geophysical Union (WDCC).
- The initial work plan will focus on engagement with identified data centres, members of WDS, holders of the Data Seal of Approval, ICPSR members, and others. Should funding be available from other sources, the case studies will be expanded to include a larger number of data centres. In particular, a parallel activity would look at the cost recovery models for the European Research Infrastructure Consortia[10] (ERICs).

WG Members[11]

- Ingrid Dillo (DANS, NL) [**CHAIR**]
- Sarah Callaghan (BADC, UK)
- Simon Hodson (CODATA)
- Jared Lyle (ICPSR, US) tbc.
- Barbara Sierman (KB, NL)
- Frank Toussaint (DKRZ, Germany)
- Mark Thorley (NERC, UK, observer)
- Kim Finney (AAD, Australia)
- Anita de Waard (Elsevier, US)
- Eva Zanzerkia (NSF/GEO, US)
- Mikael Karstensen Elbæk (OpenAIREplus)

References

All of the working groups in the Publishing Data Interest Group have a common bibliography[12] in which publications relevant for this particular group are marked correspondingly.

[10] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric

[11] This list will be expanded, especially with representatives from data centers. There are a number of ideas of people who could be contacts (representatives of ADS, UKDA, ICPSR, CCDC, DCC, 4C).

[12] Bibliography: http://goo.gl/wA1G27