Interim Report on Data Inventories and DMPs

Summary

As a core component of the IDRC Pilot on Data Sharing each of the seven participating projects developed a data inventory and used existing templates or services to prepare a Data Management Plan. The process of developing the Inventories and Plans has raised a range of issues for both the design of the planning process, the support necessary for projects, local capacity for delivering on data sharing and how this interacts with local concerns and attitudes to the control over and sharing of data.

- For most participants the concrete process of the Data Inventory, focussed on identifying specific outputs was more helpful than the more abstract question posed by existing DMP templates and tools.
- Online tools for DMP preparation were not appropriate in a number of settings due to local network capacity and reliability and modes of working.
- All projects expressed benefits from the process of thinking more clearly about the data they were generating and how to manage it.
- Most projects identified a wish to be able to provide a sharing platform for access to data, but a substantial proportion did not have the internal capacity to deliver on this.
- All projects in different ways expressed a desire for control over the process of sharing. There is a mismatch between the desire for control and the aspirations of the IDRC policy for open data which will need to be addressed directly.
- The motivations for control differ between projects but a common theme is the desire (and perceived desire of the funder) for information on who is accessing the data and for what purposes.

Introduction

A core part of the IDRC pilot program on Data Sharing was to provide the participating projects with the opportunity to define their own agenda and scope for data sharing. The design of the pilot included a process of developing an initial Data Inventory that identified expected data products, defined their formats and size, and investigated the issues that data sharing and management would raise. This was followed by the preparation of a Data Management Plan, with the preference being the use of the Portage DMP Assistant Tool. The DMP Assistant Tool is an internationalisation of the Digital Curation DMP Online service that was selected because it is bilingual (English/French). The default set of questions was used.

The process of preparing both inventories and plans was slower than planned and required more chasing of participants than was expected. In some ways these reinforces the value of requiring planning as part of the submission process for grant proposals but, as noted in our previous survey, this raises issues of whether the planning process becomes viewed as a purely administrative requirement. All participating projects noted that the process was valuable.

The Data Inventory and Data Management Planning Process

The Data Inventory was based on a table template. The template was introduced to participants as part of a workshop after a discussion that was intended to expand their understanding of what might be classified as research data. Most participants found the template helpful and stated that it helped to focus them on the scope of objects they needed to consider.

Comment [BP1]: It will be interesting to discuss how this can be overcome.

Comment [RS2]: This internal capacity issue is very important. Did any of the teams manage to build internal data management capacity beyond the person(s) who attended the meetings and were the main contact for the project? If so, what capacities were developed, lessons learned, etc. If not, should this be considered as a weakness of the pilot project?

Comment [SG3R2]: Agree. LASDEL had discussions with all members who agreed to support the pilot activities and identified key concerns. Data management was integrated into job description of research coordinator; an assessment was done of IT capacity and needs. Thus foundation was laid; but lack of capacity to implement recommendation.

Comment [RS4]: What level of control?

Comment [MI5]: I think this issue and the one below may be clarified at the next workshop – I have the impression that the aspirations of IDRC may have been misunderstood.

Comment [BP6R5]: Yes, I agree.

Comment [RS7]: I wonder how many of the projects had this perception and how they got this impression?

Comment [BP8]: This would be helpful to determine if the benefits of open data are universal or greater for select groups of researchers i.e. those from the developed world.

Comment [RS9]: Was the delay caused by the lack of responsiveness or simply the lack of the required skills to complete the tasks as instructed?

Comment [SG10R9]: In LASDEL's case it was 2 things: skills – and Pascal provided tremendous support in clarifying and reviewing work done. It also was issue of time required – inventory and DMP are 'invisible work' added to already overflowing commitments.....

Comment [11]: Maybe useful to include this in an appendix.

By contrast the introduction of the DMP tool was less successful. Participants found the questions posed abstract and not easy to understand in context. In some cases the questions did not seem well posed for their project. In the case of one participant experienced with handling DMPs the general scope of questions was seen as valuable because they could be addressed in a way that was shaped by the specific context of the project but most other participants struggled with the scope and intent of the questions.

Comment [MI12]: Is this due to a need to clarify the concepts for better understanding or is the tool not appropriate for different contexts (then a need to develop new ones or improve existing ones)?

Interviews with the project participants reinforced the need for support in contextualising the process of Data Management Planning and therefore underline the conclusions of the review that the provision of support by funders is critical in implementing a requirement for DMPs. Current services focus on the capability of a given funder or institution to provide a templated set of questions but do not support a contextual or dynamic set of questions that adapt to the specifics of a given project. There is a substantial user experience challenge in developing systems that are sufficiently flexible to support a wide range of projects yet able to guide those new to Data Management Planning through the process.

Comment [13]: Maybe add a short section earlier regarding the methodology of the study, including interview questions?

Comment [BP14]: Possibly a significant challenge considering the multi-disciplinary nature of the research funded by IDRC.

The use of an online tool was problematic for several participants. This was due in some cases to network capacity or reliability (with both being separate issues) but also due to preferred patterns of working. The preferred alternative in all cases was to download the set of questions as a Word Document and to work locally with that. This suggests that where the preference is for online provision in the browser that good provision of offline functionality and or local cacheing (as available for instance in Wordpress and GoogleDocs) will be crucial.

All projects reported greater clarity and understanding of the outputs of their project and a sense of being more in control of the process of dissemination as a result of planning. Most projects also reported a concern with the increase in scope of the outputs as a result of this process. In several cases a positive decision was taken to rule some outputs as out of scope for the purposes of the Data Management Planning process.

Comment [RS15]: Again, was this because of time limitation, lack of necessary skills (or both) or a perception that some outputs are not important?

Planning for sharing

Each project, in its own way expressed a strong desire to retain control over access to project outputs and in most cases explicitly rejected third party repositories as an acceptable path to sharing. In the case of the Virtual Herbarium this desire for control is deferred upstream due data providers, but remains an important component of enabling data sharing. For the Tobacco Economics Data project the need for control was stated as being required to monitor and report on users. In the case of this project a robust data sharing infrastructure is already in place, as well as experience and history of justifying its further funding. For Vietnam much of the infrastructure is in place as a private environment, and issues arise of how to open it up technically.

In the case of the other projects (HarassMap, Derechos Digitales, Natural Justice and Niger) a need for control over data access was expressed as either arising from issues of data privacy or ethical concerns. In each of these cases a fear of undesirable and uncontrollable re-use was expressed. In some cases (HarassMap) there was also a concern to identify how best to work with downstream users. In the case of Natural Justice there is a fundamental issue of the conception of digital materials from the project *as* data i.e. as transportable, de-contextualised information.

Comment [RS16]: This is a very important issue and will have huge implication on resources, usability and sustainability. My concern is that if solutions are found to deal with privacy and security, will the projects still want to retain control?

Comment [SG17R16]: I know LASDEL is also concerned about others plagiarizing their data without giving credit. This issue has also been raised in the literature – who will most likely benefit from the open data? It could be the bigger institutions in the north.

In each of these cases the decision was made to provide data outputs on a portal to be developed by the project organisation. For HarassMap and Derechos Digitales this is a new effort to be developed from scratch. These organisations do not currently have the capacity or expertise internally to tackle server provision and maintenance and are developing this internally. For both Natural Justice and Niger server provision is not easily feasible. In Niger the focus is initially being made on enabling sharing amongst researchers. Similarly for DD the initial focus is on sharing amongst the project team.

The desire for control is seen in many aspects across research and runs counter to the data sharing policies of most funders, which focus on open sharing. Ethical and management concerns for researchers are the basis of the justification for controlling *all* data whereas funder policies tend to emphasise the release of data by default with protections for specific subsets. This policy implementation issue play out in many areas but will need to be addressed with support and education if IDRC is to pursue an Open Data requirement.

The tendency to focus on local provision for data sharing and access control raises capacity and efficiency issues. It is unclear whether most groups have the technical expertise to develop, run and manage a secure environment of the form they envision, or how this would be sustained in the long term. Most are substantially underestimating the cost and effort involved in managing a stable and secure platform. Centralised provision for archiving and sharing could address this issue and also offer an opportunity for guidance, education and standardisation of any access restrictions. This tension will need to be addressed for successful policy implementation.

Comment [SG18]: Although test data was submitted to a repository, Data First, who have not responded in terms of whether data is ready to share.

Comment [SG19]: My understanding is that it also relates to who will have more immediate access to the data to do follow up research. Often much more can be done with the data than can be completed within the framework of a single project. Many grantees, however, lack core funding and are continually having to seek new project funding – which does not include the funds to continue mining previously connected data. Thus the question – who will ultimately benefit from open data – needs to be put on the table. Funders like IDRC could perhaps provide ‘additional synergy grants’ that allow grantees to do some follow up work on data before putting it in the public domain? The issue of power relations across institutions is important.

Comment [RS20]: Even if they do, the other question is how expensive this could become (e.g., in the case of HM to start from scratch)??

Comment [SG21R20]: Agree completely – for example LASDEL needs an intranet system (server etc) and faster connection bother of which are expensive; difficult when dependent on varying project funding and varying levels of indirect costs. Would IDRC include these aspects under a ‘data management capacity’ budget line?

Comment [22]: So one clear emerging lesson is that data sharing is more of a socio-political issue and not a technical one. This make policy provision far more challenging.

Comment [RS23R22]: I absolutely agree.

Comment [SG24R22]: Well put. Me too!