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.

Comment [LW1]: I'm not sure I understand what you mean here. Do you mean that projects don't want to have strictly Open Data, as this precludes registering of researchers who wish to use the data? Open = Free and online, at the lowest level of aggregation possible, for any purpose. However, I don't think that requiring users to register and state their data purpose is at odds with an Open Data approach, as long as the data is available for any purpose, that is, as long as no listed purpose prevents access. I think this is not a form of mediated access, but rather statistics keeping. Funded organisations may need to keep a record of how valuable their data is, and how it is used, to report to funders.

Introduction

The IDRC Data Sharing Pilot is based on following the development and implementation of a data management process by seven participating volunteer projects. This involved supporting the projects to first inventory the forms of data they expected to produce using a provided template (see Appendix A) and then to develop a Data Management Plan (DMP), using the DMPAssistant Platform developed by Portage, Canada. After the preparation of the DMPs a semistructured interview was carried out via Skype (interview prompt is in Appendix B). Analysis of the DMPs prompted by comments from the interviews forms the basis of this report.

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.

Comment [LW2]: This may be the case. But funders are now wising up, particularly where projects request funds to undertake secondary analysis on country data. Funders have begun to check to see if the data presented as evidence for research is available and appropriate for the research.

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.

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.

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.

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.

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

Comment [LW3]: We found that simply by listing what we planned to collect, we came across misunderstandings between project researchers who will collect the data, and DataFirst staff who will prepare, preserve and share the data. The misunderstandings were largely around the economics focus of the data. DataFirst expected the Project to collect tobacco production and consumption data only, and needed to be apprised of the fact that the Project, as the Economics of Tobacco Control project, has an economics focus. Once this was made apparent, project partners were able to line up our data ambitions.

Comment [LW4]: Because ours is a data-focused project, and because as technical partners we already work in the area of data management, we found the DMP equally useful. This was especially pertinent when we took the plan back to unpack with Project researchers.

Comment [LW5]: RDMPs deal with questions which are vital to any research with an Open Data agenda. We immediately discovered that the Project PI was, in fact, not able to share the data she was using as evidence for her research. This was because her previous employer had purchased this under license from a commercial supplier. It is doubtful to us if she even has the right to use the data for her current research. RDMPs immediately focus on these kinds of ethics in reuse issues. It is these issues that need to be addressed by funders developing an Open Data agenda.

Comment [LW6]: Very true. Researchers often treat these with contempt, and agree to manage and share data to obtain funding, but renege on this, often without sanction from funding agencies.

Comment [LW7]: This is definitely something that needs to be taken into account with LMIC participants from organisations without good internet access.

Comment [LW8]: So the RDMP process helped to focus projects where aims were rather vague. You can see why this would be valuable to funders before they commit funds to a project.

Comment [LW9]: Do you mean that projects don't want to have strictly Open Data, as this precludes registering of researchers who wish to use the data? Open = Free and online, at the lowest level of aggregation possible, for any purpose. However, I don't think that requiring users to register and state their data purpose is at odds with an Open Data approach, as long as the data is available for any purpose, that is, as long as no listed purpose prevents access. In that way we are not controlling what is done with the data, and this is not a form of mediated access, but rather statistics-keeping. Funded organisations like ours need to keep a record of how valuable their data is, and how it is used, to report to funders.

conception of digital materials from the project *as* data i.e. as transportable, de-contextualised information.

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 and reinforces the lesson that data sharing is more a socio-political issue than a technical one.

Comment [LW10]: In this case, funders have the power to change a data hording culture, and they should exercise this power for better Science and public good. History has shown that researchers adapt when their funding depends on this adaptation. This can involve the carrot of RDMP funding and support, and the stick of withholding future funds for non-compliance.

Comment [LW11]: It wouldn't make sense for the IDRC to go into the data archiving business. There is already a network of archives worldwide, catering for every field. A better option is to require projects to deposit their data in any accredited (e.g. Data Seal of Approval) discipline-appropriate archive, where it will be preserved and shared professionally. Funders can provide a percentage of funding to support the costs to smaller archives, e.g. the RCUK approach.

Comment [LW12]: This is very true, but funders can change this (see my comments above)

Appendix A - Data Inventory Template

Process/Work Package	Expected Data Outputs	Description	File Formats/Size	Potential Issues
	Spreadsheets			
	Questionaires or Interview Prompts			
	Interview transcripts/forms			
	Audio/Video recordings			
	Databases			
	Research records/notebooks			
	Images			
	Other			
	Spreadsheets			
	Questionaires or Interview Prompts			
	Interview transcripts/forms			
	Audio/Video recordings			
	Databases			
	Research records/notebooks			
	Images			
	Other			
	Spreadsheets			
	Questionaires or Interview Prompts			
	Interview transcripts/forms			
	Audio/Video recordings			
	Databases			
	Research records/notebooks			
	Images			
	Other			

Appendix B - Interview Prompts

Data Management Planning and plans for data sharing

- Experiences thus far?
- Specifics of the Data Planning Process:
    - Value in publishing the DMP?
    - What was useful in the Data Inventory process?
    - What was not useful?
    - What was useful in the Data Management Planning process?
    - What was not useful?
- Reflections on the process
    - What are the current plans for data sharing?
    - Has this process changed your local practice?
    - Has it changed the way you think about collecting/managing data?
    - Thinking back to yourself and your project 6 months ago, what advice would you give that person if you could today?
- Timing and logistics
    - Would DMP have been useful had it been done at the beginning?
    - Do you use the DMP as a document throughout the project?
- What are the Next Steps for implementing the plan?
- What are the Main Issues revealed by the planning process?

Appendix C - traduction française par Google Translate

Rapport intérimaire sur les inventaires de données et PGD

Résumé

En tant que composante de base du pilote du CRDI sur le partage des données de chacun des sept projets participants ont développé un inventaire des données et utilisé des modèles ou des services existants pour préparer un plan de gestion des données. Le processus d'élaboration des inventaires et des plans a soulevé une série de questions à la fois la conception du processus de planification, le soutien nécessaire pour les projets, les capacités locales pour la prestation sur le partage de données et comment cela interagit avec les préoccupations et les attitudes locales au contrôle et le partage des données.

- Pour la plupart des participants le processus concret de l'inventaire des données, axée sur l'identification des produits spécifiques était plus utile que la question plus abstraite posée par des modèles et des outils de DMP existants.
- Les outils en ligne pour la préparation DMP ne sont pas appropriés dans un certain nombre de paramètres en raison de la capacité du réseau local et la fiabilité et les modes de travail.
- Tous les projets ont exprimé des avantages du processus de réflexion plus clairement sur les données qu'ils génèrent et comment le gérer.
- La plupart des projets identifiés un désir d'être en mesure de fournir une plate-forme de partage d'accès aux données, mais une proportion importante n'a pas eu la capacité interne pour fournir à ce sujet.
- Tous les projets de différentes manières exprimé le désir de contrôle sur le processus de partage. Il existe un décalage entre le désir de contrôle et les aspirations de la politique du CRDI pour les données ouvertes qui devront être adressées directement.
- Les motivations pour le contrôle diffèrent entre les projets, mais un thème commun est le désir (et le désir perçu du bailleur de fonds) pour obtenir des renseignements sur les personnes qui l'accès aux données et à quelles fins.

Introduction

Le CRDI partage de données pilote est basé sur la suite du développement et de la mise en œuvre d'un processus de gestion des données de sept projets bénévoles participants. Cela a impliqué l'appui des projets de premier inventaire des formes de données qu'ils devraient produire en utilisant un modèle fourni (voir l'annexe A), puis d'élaborer un plan de gestion des données (DMP), en utilisant la plate-forme développée par DMPAssistant Portage, Canada. Après la préparation de la PGD une entrevue semi-structurée a été réalisée via Skype (invite interview est à l'annexe B). L'analyse des PGD suscitées par les commentaires des interviews constitue la base de ce rapport.

Une partie essentielle du programme pilote du CRDI sur le partage de données était de fournir aux participants des projets avec la possibilité de définir leur propre agenda et la portée du partage des données. La conception du projet pilote comprenait un processus d'élaboration d'un inventaire de données initial qui a identifié les produits de données attendus, défini leurs formats et de la taille, et a enquêté sur les questions que le partage de données et de gestion soulèveraient. Ceci a été suivi par la préparation d'un plan de gestion des données, la préférence étant l'utilisation de l'outil adjoint Portage DMP. L'outil adjoint DMP est une internationalisation du service curation numérique DMP en ligne qui a été choisi parce qu'il est bilingue (anglais / français). La série de questions par défaut a été utilisé.

Le processus de préparation des deux inventaires et des plans a été plus lente que prévu et a nécessité plus ciselure des participants que prévu. À certains égards, ces renforce la valeur d'exiger la planification dans le cadre du processus de soumission des propositions de subventions, mais, comme indiqué dans notre précédente enquête, cela soulève des questions de savoir si le processus de planification devient considérée comme une exigence purement administrative. Tous les projets participants ont noté que le processus a été précieux.

L'inventaire des données et des processus de planification de gestion des données

L'inventaire des données a été basé sur un modèle de table. Le modèle a été présenté aux participants dans le cadre d'un atelier après une discussion qui visait à élargir leur compréhension de ce qui pourrait être classé comme données de recherche. La plupart des participants ont trouvé le modèle utile et a déclaré qu'il a aidé à les concentrer sur la portée des objets dont ils ont besoin à considérer.

En revanche, l'introduction de l'outil DMP a moins bien réussi. Les participants ont trouvé les questions posées abstrait et pas facile à comprendre dans le contexte. Dans certains cas, les questions ne semblent pas bien posé pour leur projet. Dans le cas d'un participant expérimenté avec la manipulation DMPS la portée générale de questions a été considéré comme précieux parce qu'ils pourraient être traités d'une manière qui a été façonné par le contexte spécifique du projet, mais la plupart des autres participants aux prises avec la portée et à l'intention des questions .

Entrevues avec les participants au projet ont renforcé la nécessité d'un soutien en contextualisant le processus de planification de la gestion des données et donc souligner les conclusions de l'examen que la fourniture d'un appui des bailleurs de fonds est essentiel pour mettre en œuvre une exigence pour PGD. Les services actuels se concentrent sur la capacité d'un bailleur de fonds ou une institution donnée pour fournir un ensemble modélisée de questions, mais ne prennent pas en charge un ensemble contextuel ou dynamique de questions qui s'adaptent aux spécificités d'un projet donné. Il y a un défi de l'expérience utilisateur importante dans les systèmes qui sont suffisamment souples pour soutenir un large éventail de projets encore en mesure de guider les nouveaux dans le processus de planification de la gestion des données en développement.

L'utilisation d'un outil en ligne a été problématique pour plusieurs participants. Cela est dû dans certains cas, la capacité du réseau ou la fiabilité (avec les deux questions distinctes étant), mais aussi en raison de motifs préférés de travail. L'alternative préférée dans tous les cas était de télécharger l'ensemble des questions comme un document Word et de travailler localement avec cela. Cela donne à penser que lorsque la préférence est pour la fourniture en ligne dans le navigateur que la bonne disposition des fonctionnalités hors ligne et ou un cache local (comme disponible par exemple dans Wordpress et GoogleDocs) sera cruciale.

Tous les projets ont indiqué une plus grande clarté et la compréhension des résultats de leur projet et le sentiment d'être plus en contrôle du processus de diffusion à la suite de la planification. La plupart des projets ont également signalé une préoccupation avec l'augmentation de la portée des sorties à la suite de ce processus. Dans plusieurs cas, une décision positive a été prise à la règle des sorties comme hors de portée pour les besoins du processus de planification de la gestion des données.

Planification pour le partage

Chaque projet, à sa manière a exprimé un fort désir de conserver le contrôle sur l'accès aux résultats et dans la plupart des cas du projet rejeté explicitement référentiels tiers comme un chemin acceptable pour le partage. Dans le cas de l'Herbier virtuel ce désir de contrôle est reporté en amont les fournisseurs de données dues, mais demeure un élément important de

permettre le partage des données. Pour le projet Tobacco Economie Données de la nécessité d'un contrôle a été déclaré comme étant nécessaire pour surveiller et faire rapport sur les utilisateurs. Dans le cas de ce projet d'une infrastructure de partage de données robuste est déjà en place, ainsi que l'expérience et de l'histoire de justifier davantage son financement. Pour le Vietnam une grande partie de l'infrastructure est en place comme un environnement privé, et des questions se posent sur la façon de l'ouvrir sur le plan technique.

Dans le cas des autres projets (HarassMap, Derechos Digitales, de la justice naturelle et Niger) un besoin de contrôle sur l'accès aux données a été exprimée soit résultant de problèmes de confidentialité des données ou des préoccupations éthiques. Dans chacun de ces cas, la crainte d'une réutilisation non souhaitable et incontrôlable a été exprimé. Dans certains cas (HarassMap) il y avait aussi une préoccupation pour identifier la meilleure façon de travailler avec les utilisateurs en aval. Dans le cas de la justice naturelle, il est une question fondamentale de la conception de matériaux numériques du projet en tant que données à savoir que transportable, des informations de-contextualisée.

Dans chacun de ces cas, la décision a été prise pour fournir des sorties de données sur un portail à développer par l'organisation du projet. Pour HarassMap et Derechos Digitales ceci est un nouvel effort pour être développé à partir de zéro. Ces organisations ne sont pas actuellement la capacité ou l'expertise en interne pour lutter contre la fourniture de serveurs et d'entretien et se développent de manière interne. Pour les deux la justice naturelle et la fourniture du serveur Niger est pas facilement réalisable. Au Niger, l'accent est d'abord faite à permettre le partage entre les chercheurs. De même pour DD l'objectif initial est sur le partage entre l'équipe du projet.

Le désir de contrôle est considéré dans de nombreux aspects à travers la recherche et va à l'encontre des politiques de partage des données de la plupart des bailleurs de fonds, qui mettent l'accent sur le partage ouvert. Les préoccupations éthiques et de gestion pour les chercheurs sont à la base de la justification pour contrôler toutes les données alors que les politiques des bailleurs de fonds ont tendance à mettre l'accent sur la diffusion des données par défaut avec des protections pour les sous-ensembles spécifiques. Cette question de la mise en œuvre des politiques jouent dans de nombreux domaines, mais devra être abordée avec le soutien et l'éducation si le CRDI est de poursuivre une exigence Open Data.

La tendance à se concentrer sur la fourniture locale pour le partage de données et de contrôle d'accès pose des problèmes de capacité et d'efficacité. Il est difficile de savoir si la plupart des groupes ont l'expertise technique pour développer, exécuter et gérer un environnement sécurisé de la forme qu'ils envisagent, ou comment cela pourrait être maintenu à long terme. La plupart sont sous-estiment considérablement le coût et les efforts nécessaires à la gestion d'une plate-forme stable et sécurisée. fourniture centralisée pour l'archivage et le partage pourrait résoudre ce problème et offrir également une opportunité pour l'orientation, l'éducation et la normalisation des restrictions d'accès. Cette tension devra être adressée pour la mise en œuvre réussie des politiques et renforce la leçon que le partage des données est plus une question socio-politique que technique.