|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Meeting title:** | | DataPool Steering Group Meeting | | |
| **Date:** | | 31st May 2012 | **Time:** | 12:30-15:30 |
| **Location:** | | John Roberts Room, Staff Social Centre (B38), Highfield Campus, University of Southampton | | |
| **Present:** | | Dr M. Brown (chair); Dr L Carr; Prof S Cox; Dr G. Earl; Dr P. Hancock; Prof P. Nelson (Pro VC Research); M. Ployaert; Dr H. Snaith; Prof A. Wheeler (Provost and DVC); W. White.  L. Corti (Associate Director, UK Data Archive, University of Essex); G. Pryor (Associate Director, Digital Curation Centre, Edinburgh); S. Rumsey (Digital Collections Development Manager, Bodleian Libraries, University of Oxford)  Project Managers: D. Byatt; Dr S. Hitchcock | | |
| **Apologies:** | | Prof J Frey | | |
|  | |  | | |
|  | **Agenda item** | | | |
|  | **Research Data and DataPool : an update**  Mark Brown (MLB) summarised the key points from Paper 1 DataPool Update Report. There were two phases to DataPool – Phase 1 Infrastructure and Phase 2 Institutional rollout. Challenge is to move beyond a discipline focussed model towards an institutional one. This will require cultural change. Driving factors includes Funding Council requirements and government directives on access to research data and results. DataPool also needs to be a researcher orientated project. | | | |
|  | **Research Data Management Policy and Guidance** (W. White)  The University Research Data Management Policy was developed in an international context, learning from work in Australia and the US and issues arising from key funder policies such as the EPSRC and EU. Working through the Research and Enterprise Advisory Group and designated Faculty data contacts there was discussions and comments on various drafts. The process was been seen as an iterative one, similar to the Monash University approach. Senate approved the policy in February with the expectation that it would be re-presented in a year or so. Guidance development will also be iterative.  A key challenge is business modelling to support policy and this currently being investigated.  EPSRC requires a metadata catalogue and there are benefits in this but the enticement for researchers to do this lies elsewhere perhaps in the storage.  Key practical question to be addressed - who is responsible for what? If keeping data for 10 years and the individual leaves the institution - how is this handled? Is this the responsibility of the individual researcher, the PI, the institution and/or external repository?  Significance testing is a key element – IDMB survey suggested that researcher wanted to keep forever. Responsibility for this lies with the PI, so what does that mean for the PI?  Suggestion that the experience of PI’s differs and that help would be required, perhaps in the form of a checklist.  In response to a question about existing support backups and proposed infrastructure. Compliance by 2015, is recognised as very important, but need to understand more clearly what is needed, how much is needed, how is it best provided? Important to take a considered view so that provision will be fit for purpose.  Issue raised over significance as this is something yet to be tested. There is a link between the human created metadata and significance. Human creation of metadata is a very resource intensive operation; the requirement to create it may drive the definition of significance. Need to find a way to make it practical and contain costs – automation one area.  Need to look at way to get money from Funders; does seem to be some buy-in from Research councils, but need to address how this is presented to avoid producing “expensive” proposals. Costs need to be more visible, not hidden in FEC figures where this is permitted. Suggestion that NERC allows for 6% of their research programme budget to go on data management so expect that to be translated into projects. EU projects can have requirement to deposit in specific repositories.  Need to address cost and place in workflow as well as making clear the difference between working data space and archive as this is different spaces, requiring a different service/support.  The meeting recognised that there were some funders requiring to hold data and assign the DOI and some who only need to know where it is held, but are not prescriptive on the where.  Guidance – starting to develop, engage with researchers and link with Research Council requirements. Raise awareness and be sign posts. There was positive feedback on the sample guidance sheets circulated and were of a good length.  **Action:**   * Investigate ways in assisting PI’s develop criteria to identify significant data * Steering Group to advise on additional guidance topics | | | |
|  | **Infrastructure: Overview**  **SharePoint (P. Hancock)**  Description of the purpose of the SharePoint as a workflow tool was given. It would allow for deposit into a large archive Institutional repository or an external one with the aim of matching up the fields required for that repository. Have begun to develop a metadata front end.  SharePoint was proposed because of the flexibility of what it can offer – a working collaborative environment that can be self-managed, a means of storing material during the life of a project and over the longer term, a custodial location, a vehicle for managing material. On the other hand ePrints has more of a focus on end deposit.  SharePoint has the potential for closer desktop integration compared with having to work with interoperability protocols.  SharePoint has certain features that will allow pre-loading of terms or schema. It also has the facility to allow external collaboration and public visibility. Could help with avoiding inappropriate use of other services such as Dropbox that raise issues relating to IP.  **ePrints (L. Carr)**  In repository world science dominated by really big data, but the real challenge is to keep the smaller scale projects in focus – many more of these and more complex problems. . ePrints is well placed to deal with these. Big data is often automated  Important not to develop a system that imposes schema or taxonomy. Need a balance of right level of activity and support provided. Data-oriented ePrints record in development and a ePrints bazaar package now available. Will deal with 3 types of document:   1. Data 2. Additional metadata 3. README   Won’t require data itself to be download, this can still be stored elsewhere. Bulk upload would be dealt with through SWORD. Links here to other JISC work including things developed via DepositMO and DepositMore. Although outside the remit of this project, there is a need to develop some SWORD based ingest methods.  In response to a question on mandatory fields, it was stated that there would be a minimum set of core fields with the requirement for other fields left flexible in order to accommodate all the various types of data.  Level of metadata has to be sufficient to enable the data to be “picked up” so balance between what could and what needs to be done. UK Data Archive grade submissions and apply added value as appropriate. Suggestions that there is an institutional role in evaluating the data, but not easy. Doesn’t have to be immediate access. Can be “it’s here and we can get it if you want it”. Often people want to ask about the data not simply download it.  Repository needs to provide additional services, not just storage, for example, adding consistent metadata. Coming from the UK Data Archive experience it is essential that data covering sensitive areas is checked prior to release. Also needs permissions to be checked on the re-use of data obtained under license where there might be restrictions on sharing. Might be possible to link this to ethics approval procedures in the future, but will need training for ethics team.  Need to recognise that we are still judging things on existing procedures that were set up 10 +years ago and these may not go into the future. There are new influences that will mean people will have different attitudes and approaches.  There is evidence from EPrints that public citation does raise profiles.  **Storage (P. Hancock**)  Recognise that there is a need to manage data throughout the full life cycle. Storage is not just about a place to store the data, but using a managed Filestore is a way to guarantee accessibility for the researcher and in the longer-term make it available to others. Had hoped as part of this project to be able to offer deposit from inception all the way through to the end of the project, enabling easier curation and preservation. Recognise need to provide a system/strategy to secure data using a location that enables preservation, emphasising that iSolutions storage is much more than the easy and cheap quick fix solution of buying a drive. Want to move to a position where we are not only holding active data but also store data safely and effectively for the longer-term. It does come down to investment and a need for training of future researchers in good data handling practices.  Important to tailor resources but not so much that it is limiting provision. It is difficult to predict what will be significant in the future. May want to have some showcases, but not restrict only to those areas. Want to pilot and experiment, to allow success to pull through others to success, to align with researcher interest.  Need to acknowledge that there are different requirements in different disciplines and the funding also differs. Some estimate that 20% need 80% of the storage so can’t develop an average by researcher figure. Also need to recognise frequency or infrequency of data re-use, possibly 99% is for latest material and the speed of retrieval. Important to develop some idea of where we are as an institution on who needs what amount.  It is important that any solution is flexible.  **Service Model (W. White)**  Challenge is how we knit together the research community with the services. Researchers don’t care who answers, just want the answer. Would like to provide a one stop shop style solution to support.  Experience from the UKDA suggests that all enquiries should feed into FAQ’s. There is a key triangle of Library, IT and Research Support. The UKDA and DCC have central helpdesks  There is some interest in drawing together data management, environmental impact statements and REF as a way to improve engagement and value. This will enable what people are already doing to be recognised.  Rollout will start with individuals using their day to day contacts to raise awareness of the availability of data management support. There will be lots of routes in to support and to engage with the iterative approach to its development using existing contacts.  Discussion took place concerning how best to raise awareness of researchers responsibility to manage their data. It might be something that is included in employment contracts but it is about making that relevant to everyday practice. There is a presumption that everyone knows this is what they should be doing. | | | |
|  | **Training Case Studies (W. White; Simon Cox**)  Simon Cox talked about the reasons behind the development of an “Introduction to Research Data” (P4a) by Mark Scott et al. Five case studies to hang your ideas on. Give examples of scientists who have acknowledged a value in making data available. Trial group (1st year postgrads) thought it useful and made them think. Making it available for future scientists.  Want to use the document as a framework for developing this more widely in the University by working on additional cases studies from different disciplines.  Although focussed on postgraduates it was thought that it could be used with other groups.  Need to approach training at different levels – the early researcher who asks “what do I do with my data” and the supervisor who has to answer the questions.  Not just about technology, but about getting people to engage. Incremental change will facilitate cultural change. Discipline good practice examples such as Archaeology and the Digital Economy University Strategic Research Group can help facilitate this.  Case Study 4 Sharing data safely  In developing training resources discipline differences in willingness to share and awareness of requirement to share need to be addressed.  Assigning metadata will require a variety of approaches to accommodate the different types of data, the level required, but it cannot become an onerous process. Key is adding sufficient to facilitate visibility. Balance between what can be automated and what perhaps an experienced PI should add manually that would better describe the content. There are data extraction tools available that can help with this.  Case Study 2 Archiving project tweets, blogs and websites  There was a brief description of the social media case study based on a conference that took place at Easter where people were encourage to use social media, Twitter in particular. It is looking at the method of capturing and archiving tweets, investigating the ethical and legal issues. There will also be the need to look at what data is kept and what can be made available. | | | |
|  | **Dissemination**  There are two aspects to this. Firstly it is a case of facilitating good practice. This may be best done in with top down support and requires the storage and infrastructure to be in place. Secondly it is about sharing of data and this may be best coming from researchers at ground level where impact can be most clearly demonstrated. | | | |
|  | Next Steps  MLB saw the following as part of the longer term processes to be investigated:   * Review of type of storage * Modelling of significance * Data registry * Case studies with discipline interest | | | |
|  | AOB  Documents were commended and request that the material could be made available. | | | |
|  | Date of Next Meeting 12th November 2012 | | | |