Data Management Plans and Data Centers

Denise DiPersio, Christopher Cieri, Daniel Jaquette
Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
{dipersio, ccieri, jaquette}@ldc.upenn.edu

Abstract

Data management plans, data sharing plans and the like are now required by funders worldwide as part of research proposals. Concerned with promoting the notion of open scientific data, funders view such plans as the framework for satisfying the generally accepted requirements for data generated in funded research projects, among them that it be accessible, usable, standardized to the degree possible, secure and stable. This paper examines the origins of data management plans, their requirements and issues they raise for data centers and HLT resource development in general.

Keywords: language resources, data distribution, data management plans

1. Introduction

Data Management Plans (DMPs) and their equivalents have become familiar to researchers developing language resources over the past several years. Many funding agencies around the world now require that research proposals include a specific section detailing how resulting data will be created, shared and maintained. For instance, the US National Science Foundation's (NSF) Data Sharing Policy states: *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.*[1] It is the US government's intention to broaden such sharing mandates by requiring federal agencies of a certain size to ensure that direct results of federal funding are publicly available.[2] Funders' guidelines on DMPs encourage or even require behavior that is not universal in the research community, stirring debate over the best ways to comply. In some instances, requirements are deliberately unclear, leaving it to the community to reach a resolution. For example, DMPs may suggest or require that language resources (LRs) are distributed at some minimal cost. Knowing though that data management costs cannot be completely eliminated, data centers must now identify new funding strategies. This paper describes different sets of DMP requirements principally in the United States, discusses issues relating to their implementation and assesses their potential impact on the availability of LRs.

2. Data Distribution in the Pre-DMP World

The majority of LRs currently available were created before or outside of the influence of DMPs though the motivations for DMPs – broad and affordable access to digital data – are long-standing. Data centers have approached this goal in multiple ways: low cost or subsidized corpora, limited free distributions under an author-pays or funder-pays model and grants in data subsidized by the community.

For example, LDC distributes several data sets at no cost either because the funding agency paid in advance for a number of copies to be distributed or because the author and LDC agreed to share costs: American National Corpus (Reppen, Ide, Suderman, 2005), AQUAINT (Graff, 2002), FORM Kinematic Gesture (Martell et al., 2004), Grassfields Bantu Fieldwork (Bird, 2003), Proposition Bank (Palmer et al., 2004) and Santa Barbara Corpus of Spoken American English (DuBois et al., 2000). Some commercial organizations subsidized the distribution of their contributions; Web 1T 5-gram Version 1 (Brants, Franz, 2006) is an example. Finally, LDC organizes a semi-annual competition for data scholarships. Applicants submit a research plan and an adviser's letter expressing confidence in the project and lack of funds. The program is subsidized by LDC membership fees. To date, LDC has awarded 64 grants valued over US$175,000.

3. Perspectives on DMPs

Open access initiatives in the United States concerning research data are not new. In 2007, the National Institutes of Health established PubMed Central, the online archive containing electronic copies of peer-reviewed manuscripts resulting from funded work. The Public Access Policy Forum, launched by the White House Office of Science and Technology in 2009, solicited public views on access to funded research results. Although the primary focus was still journal articles, it was acknowledged that research results include data sets as well. The result was a policy statement that digital scientific data supported by federal

[1] NSF Data Sharing Policy, http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_6.jsp#VID4

[2] Executive Office of the President, Office of Science and Technology Policy, Memorandum (February 22, 2013).

funding should be publicly accessible at no cost to users.[3] NSF institutionalized its long-standing sharing policy by requiring data management plans for all proposals in 2011. Those sentiments were echoed internationally in 2013 at a meeting of the G8 Science Ministers to discuss global challenges to research, including access to open scientific data. They concluded: "*publicly funded scientific research data should be open*", and it "*should be easily discoverable, accessible, intelligible, useable and wherever possible interoperable to specific quality standards.*"[4] Implementation of those ideas can be seen in current European Union (EU) and United Kingdom (UK) research programs.

The EU Framework Programme for Research and Innovation, HORIZON 2020, includes guidelines for open access to research results. Open access is defined as "*the practice of providing on-line access to scientific information that is free of charge to the end-user and that is re-usable.*"[5] Scientific information includes "*research data*", that is, "*data underlying publications, curated data and/or raw data*."[6]

UK government funders take varying approaches, all of which include some sort of data management or data sharing plan in which researchers explain how project data will be managed. Plans cover data types and formats, standards, metadata, preservation, security and sharing.[7] Similarly, the National Research Foundation (NRF) of South Africa encourages its community to develop open access policies and establish open access repositories to promote innovation and to spread knowledge, among other things. As of March 2015, papers describing funded research should be deposited in an institutional repository, and data developed in a program should be archived in an accredited Open Access repository with a Digital Object Identifier.[8] Existing repositories, like the Language Resource Management Agency, are already taking on that role.[9]

The Australian Research Council likewise has an open access policy for research findings. Publications must be deposited in an open access repository. Researchers are encouraged to deposit data generated from a project as well, since "*data management is an important part of the responsible conduct of research.*"[10]

4. Components of Data Management Plans

As seen above, most governments and funding agencies agree on the essential elements of open access and how it should be achieved. Some require data management plans, while others encourage submissions of DMPs or some equivalent that explains how data in the project will be managed and made accessible. The elements of those submissions are similar as seen below and in Table 1.

4.1 US Funding Agencies

US funders rely as a starting point on the government's definition of "research data": "*the recorded factual material commonly accepted in the scientific community as necessary to validate research findings*"[11], but they leave it to the community to define the research outputs covered under a plan. DMP guidelines focus on data and academic papers more than tools, and certainly more than metadata and specifications, notwithstanding the importance of the latter to understanding, exploiting and replicating resources. For current purposes, we believe that the definition of language resources should include any assets that allow research to succeed.

NSF requires a description of the data; the hosting archive; details of access and sharing, including re-use distribution and derivatives; the metadata used; intellectual property rights; privacy or ethical issues; data, tools and documentation formats; and plans for archiving and preservation. NSF is clear that costs for data management can be covered by the project and should be included in the proposal budget.

Two of the NSF directorates most relevant to LREC, **Computer & Information Science and Engineering (CISE) and Social, Behavioral and Economic Sciences (SBE)**, provide supplemental guidance. CISE and SBE want to know how investigators will manage and disseminate project data, including plans if investigators leave their institutions or the project. CISE also suggests that researchers consider repository options and requirements including the researcher's home institution and any community-recognized repositories, and provide a contingency plan if the chosen repository becomes unavailable.

Other US agencies that have funded language-related search, such as the **Defense Advanced Research Projects Agency (DARPA), Intelligence Advanced Research Projects Activity (IARPA)** and **the Department of Homeland Security (DHS)**, do not require data management plans to be submitted with research proposals. DARPA makes research outputs available in the DARPA Open Catalog.[12] IARPA distributes some data through

[3] Executive Office of the President, Office of Science and Technology Policy, Memorandum (February 22, 2013).
[4] G8 Science Ministers Statement 13 June 2013, https://www.gov.uk/government/news/g8-science-ministers-statement.
[5] Guidelines of Open Access to Scientific Publications and Research Data in Horizon 2020, Version 16 (Dec. 2013) at 2.
[6] Ibid.
[7] Digital Curation Centre, Funders' data plan requirements, http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.
[8] Statement on OpenAccess to Research Publications from the National Research Foundation (NRF)-Funded Research, https://updatalibrarian.files.wordpress.com/2015/02/nrf_open_access_statement_19jan2015.pdf.
[9] Aims of the RMA, http://rma.nwu.ac.za/index.php/aims/.
[10] Australian Research Council, Funding for schemes under the Discovery Program for the years 2015 and 2016 at 19, http://archive.arc.gov.au/archive_files/Funded%20Research/1%20Discovery%20Programme/Discovery%20Projects/2016/Discovery_Program_2015-16_funding_rules_DP.pdf.
[11] 2 CFR 215.36(d)(i).
[12] DARPA Open Catalog, http://opencatalog.darpa.mil/.

LDC under terms that allow for broad access at no more than the incremental cost of acquiring the data. DHS also maintains a Research Catalog that contains data from various government offices, but not necessarily from funded programs.[13] The **US Office of the Director of National Intelligence (ODNI)** has created a Government Catalog of Language Resources (GCLR) as a way to both measure and improve its return on research investment. The GCLR lists resources held by government agencies as well as those hosted by external repositories. The principal audience for the GCLR seems to be government researchers though the project team has publicized its methods.[14]

The **US Department of Energy** provides comprehensive information about the contents of DMPs which are required by its Office of Science. Those include: data, metadata content, format, plans for documentation and annotation, how data will be shared and if it will be subject to restrictions; resources needed for sharing; plans for protecting any sensitive information, intellectual property rights, etc.[15]

4.2 Europe, United Kingdom Funders

Horizon 2020 includes an Open Research Data Pilot that is mandatory for certain research areas and voluntary for others. It applies to data and metadata needed to validate research results and to "other data" and "associated metadata" addressed in a DMP. The DMP must describe the data to be developed, the metadata and the standards by which it will created, details on sharing, including plans for access and restrictions on access, and a proposal for long-term preservation.[16] Participants must deposit data in a "research data repository" and provide access to users at no fee.[17]

Two UK agencies that fund language-related programs are the **Arts and Humanities Research Council (AHRC)** and the **Economic and Social Research Council (ESRC).** Both require DMPs (or Technical Plans), in the case of AHRC where digital output or technology is essential to the outcome and with respect to ESRC, for any research generating data. AHRC plans must include a description of data, formats, and size as well as information about preservation (after the project ends) and sustainability (continuing accessibility).[18] DMPs for ESRC proposals must include basic information about the data as well (volume, type, quality, formats, documentation, metadata), plans for storage, back-up, archiving, and a discussion of any issues in sharing data related to confidentiality, legal rights, etc.[19]

DMP Requirements Across Agencies and Countries

Place	Agency	DMP Required	Funding	Constraints: privacy, etc.	User Fee	Repository Provided	Scope
US	NSF	Yes	Yes	Allowed	Yes ④	No	primary data, samples, physical collections, software, models, supporting materials, journal articles, conference papers
US	DARPA	No	N/A	N/A	N/A	DARPA Open Catalog: public material from DARPA, programs; data, tools, papers	N/A
US	IARPA	No	N/A	N/A	N/A	N/A	N/A
US	Dept. of Energy	Yes	Yes	Allowed	N/A	DOE Data Explorer - DOE data collections; PAGES -- articles & manuscripts from DOE projects, Can use other repositories also	digital research data; as defined in CFR but stored digitally
US	Dept. of Homeland Security	No	N/A	N/A	N/A	Data catalog -- immigration, maritime, FEMA data; not necessarily from funded programs	N/A
EC	Horizon 2020 Framework	Yes ①	Yes	Allowed	None	No	data & metadata for validating results in publications; data & metadata generated in project
UK	Arts & Humanities Research Council	Yes ②	Yes	Allowed	None ⑤	No	activities that involve creating, gathering, collecting, processing digital information
UK	Economic & Social Research Council	Yes ③	Yes	Allowed	None	No; use responsible digital repository; ESRC Research Catalogue contains some project outputs & info about awards	research data, metadata
South Africa	National Research Foundation	No	N/A	N/A	N/A	No	funded publications & supporting data should be deposited in accredited repositories
Australia	Australian Research Council	Yes	N/A	Allowed	N/A	Australian National Data Service works with universities and other collaborators on research data infrastructure	data generated through proposed project

① Yes, in Open Research Data Pilot; optional for other program projects
② Yes, where digital output/technology essential to outcome
③ Yes, for any research generating data
④ Incremental costs allowed except for journal articles and conference papers in proposals after 01/2016
⑤ Default is none but cases for fees considered

Table 1: International DMP Requirement

[13] Department of Homeland Security, Data, https://www.dhs.gov/topic/data

[14] Return on Investment for Government Human Language Technology Systems, http://amta2012.amtaweb.org/AMTA2012Files/papers/gov-mas.pdf.

[15] US Department of Energy, Office of Science, Statement on Digital Data Management, Suggested Elements for a Data Management Plan, http://science.energy.gov/funding-opportunities/digital-data-management/suggested-elements-for-a-dmp/.

[16] Guidelines on Data Management in Horizon 2020, Version 1.0 (Dec. 2013) at 5.

[17] Horizon 2020 Guidelines of Open Access at 10, 11.

[18] AHRC, Technical Plan, http://www.ahrc.ac.uk/funding/research/researchfundingguide/applicationguidance/technicalplan/.

[19] ESRC, Data management plan: guidance for peer reviewers, http://www.esrc.ac.uk/files/about-us/policies-and-standards/data-management-plan-guidance-for-per-reviewers/.

4.3 South Africa, Australia Agencies

As mentioned above, the South African **NRF** does not mandate the submission of DMPs, but its open access policies are consistent with those in other countries that require such a plan. Some South African universities provide guidance for developing DMPs, which follow the categories seen above: data description, format, metadata, storage and back-up, security, access, sharing, preservation, and so on. [20]

The Australian National Data Service is funded by the Australian government to manage Australian research data by working with partners to build infrastructure and link data. It provides various resources including links to Australian university DMP templates.[21] Those templates address for the most part the same elements found in other DMP requirements, such as information about the data to be collected (formats, metadata, size), how it will be stored and backed-up, plans for sharing and proposals for archiving and preservation.

It should be clear from the survey above that early planning for data management is essential for developing research proposals. Yet, this is often the last piece considered on the eve of deadline, or else handled so generally in the proposal, that investigators find themselves at the project's conclusion with no repository or funding. Data centers can step in at that juncture, and they do, but planning data management strategies post-project is not optimal. It often means that data centers subsidize distribution from their general reserve funds or fees are charged to users. Researchers are typically not happy when the users pay, yet data centers cannot consistently bear distribution costs without some contribution from funders, data providers or users.

5. Implementing DMPs

DMP sharing requirements may be satisfied by access through the researcher's web site, an institutional website or a data center. Data centers can simplify compliance by exploiting existing infrastructure and processes for reviewing, storing and distributing resources over the long-term. Data center services may include: (1) pre-publication review to identify and resolve content or data integrity issues; (2) assistance in preparing comprehensible data descriptions for a wider audience; (3) improved discoverability by using persistent identifiers and by sharing resources and metadata via the data centers' own catalog and via union catalogs such as OLAC (Open Language Archives Initiative) and the Universal Catalog; (4) increased outreach through papers, conference presentation and other communications such as newsletters, social media and mailing lists; and (5) greater stability by retaining each version of a deposited corpus to support benchmarking and implementing corrections as patches or new versions rather than modifications to an existing version.

Given the number of different corpora they handle, data centers are well positioned to manage property rights, privacy and ethical concerns in accordance with standard practices in these areas. As the law, regulation and practice concerning intellectual property and the treatment of human subjects evolve and become more complex, it is in turn more challenging for a researcher trained in linguistics or human language technology to keep abreast of developments and adjust their own distribution accordingly. Data centers typically maintain staff who specialize in all areas related to the archiving and distribution of language resources.

Data centers may also work with investigators to develop realistic budgets that account for the desired level of access, data size, storage and the like. To the extent that DMPs shift the burden from a user-pays to an author-pays model, or one in which users can only bear "incremental" costs, it is in everyone's interest to find a balance that preserves sufficient funding for research activities and ensures that research data remains accessible and intact.

Given their mission, stability and broad recognition within the community, data centers may also serve as a focal point of discussions concerning standards and format and collection point for feedback regarding individual corpora. For example, LDC routinely receives feedback on the corpora it distributes, including those contributed by others; some of this feedback leads to patches or improved versions of the data. In recent years, LDC has also organized two workshops on issues of metadata needed for the analysis of variation and change in language.

6. Open Issues

The growing proliferation of DMPs requires data centers to rethink existing distribution models. Doing so reveals open issues that require consideration as well.

Intended Audience. Who is the DMP or equivalent planning document intended to serve? Possible beneficiaries include funders, government researchers or researchers working on government contracts, broader research communities, unaffiliated researchers or the public. Each of these audiences has different training, expertise and access to infrastructure. Even among different research communities the questions posed and the methodologies used will vary. For example, preparing a bilingual lexicon for use by the general public may require a different approach than making the same resource available to linguists or language technology developers. One group may require interactive search via the web with fuzzy searching to capture similar sounding forms, while another may prefer a static corpus formatted to integrate with a tool widely used by their research community, even if that format is proprietary, while the third may prefer a

[20] University of Cape Town, UCT Research Data Management Plan, http://www.lib.uct.ac.za/uct-research-data-management-plan.

[21] ANDS, Funders guidelines, http://ands.org.au/working-with-data/data-management/data-management-plans.

static corpus presented in an open, standard format for which they will develop tools.

Legacy data. How do DMPs affect the distribution of resources that pre-existed them? The answer may be that there is no effect in terms of a directive or regulation, but from a practical perspective data centers must contend with harmonizing pre- and post-DMP distribution schemes. Affected areas may include infrastructure, membership and user policies and licensing.

Cost reduction. The thrust of the DMP is to reduce or eliminate user costs for data access. It is now incumbent on data centers to re-examine fixed and variable costs, understanding that DMP funding will likely require cost-efficient and sustainable archiving and distribution arrangements. The impact of Moore's Law on storage costs, the presence of commodity storage-as-a-service and the increase in network bandwidth allow data centers to reduce some costs. However, to assume that this means archiving costs approach zero is to overlook the additional services that data centers provide. Creating and maintaining agreements that protect human subjects and intellectual property, promoting data and answering questions about data, negotiating changes to agreements with data providers and users, migrating data as necessary across instances of local storage or storage-as-a-service vendors, creating, accepting and integrating patches and restructuring data to serve emerging needs all require human effort that is not subject to Moore's Law and in fact increases in cost over time.

Trusted data repositories. Funders generally agree that research data should be made available through existing repositories already serving the community. The notion of a "trusted" repository is often mentioned in this context, but what does that mean? One approach is to consider one of the many existing certifications for data repositories, like the Data Seal of Approval (DSA) developed by the Dutch Data Archiving and Networked Services.[22] The DSA allows repositories to self-assess against a set of factors, including formats, metadata, access, preservation, infrastructure and user policies. Data centers are well advised to consider benchmarking against guidelines like these so that they can confidently represent to the community and funders their ability to implement DMPs.

Defining standards: the persistent identifier. The persistent identifier, the notion that every LR carries a permanent, unique identifier, is generally endorsed in the HLT community. LRs are identified by various means, including ISBN (LDC) LDC ID, ELRA ID and ISLRN (ELRA, LDC, RMA) and DOI (CLARIN). The identifiers differ along a number of dimensions including the defining organization, how broadly they are accepted and used, whether the identifier contains sub-parts, whether the whole (or parts) have any semantics or are arbitrary, whether they are globally unique or merely unique within a repository, how the identifiers are resolved in order to locate the resource and whether there is any cost or other impediment to assigning identifiers. Even within a single system how may one resolve the question of what the identifier identifies? Is it the corpus? The metadata? Some propose that an identifier exist for all files within a data set or that it should extend to data bytes. What does it mean, then, for a data center to say that it meets the standard for persistent identifiers? This is an example of work the community must undertake as standards assume a larger role in LR distribution under DMPs.

[22] Data Seal of Approval, http://www.datasealofapproval.org.

7. Conclusion

Data centers serving the HLT community were generally founded with the mission to provide broad access to language resources to promote data sharing and scientific progress. The thinking behind DMPs is consistent with those general principles, but their implementation poses challenges to the data center model. Those challenges are surmountable, but require collaboration between researchers and data centers to balance the needs of both and in the end, provide open, stable, readily accessible and replicable language resources.

8. Bibliographical References

AHRC, Technical Plan, http://www.ahrc.ac.uk/funding/research/researchfundingguide/applicationguidance/technicalplan/.

Aims of the RMA, http://rma.nwu.ac.za/index.php/aims/.

ANDS, Funders guidelines, http://ands.org.au/working-with-data/data-management/data-management-plans.

Australian Research Council, Funding for schemes under the Discovery Program for the years 2015 and 2016, http://archive.arc.gov.au/archive_files/Funded%20Research/1%20Discovery%20Programme/Discovery%20Projects/2016/Discovery_Program_2015-16_funding_rules_DP.pdf.

DARPA Open Catalog, http://opencatalog.darpa.mil/.

Data Seal of Approval, http://www.datasealofapproval.org.

Department of Homeland Security, Data, https://www.dhs.gov/topic/data.

Digital Curation Centre, Funders' data plan requirements, http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

ESRC, Data management plan: guidance for peer reviewers, http://www.esrc.ac.uk/files/about-us/policies-and-standards/data-management-plan-guidance-for-per-reviewers/.

Executive Office of the President, Office of Science and Technology Policy, Memorandum (February 22, 2013).

G8 Science Ministers Statement 13 June 2013, https://www.gov.uk/government/news/g8-science-ministers-statement.

Guidelines of Open Access to Scientific Publications and Research Data in Horizon 2020, Version 16 (Dec. 2013).

Guidelines on Data Management in Horizon 2020, Version 1.0 (Dec. 2013).
NSF Data Sharing Policy, http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_6.jsp#VID4
PubMed Central, http://www.ncbi.nlm.nih.gov/pmc/
Return on Investment for Government Human Language Technology Systems, http://amta2012.amtaweb.org/AMTA2012Files/papers/gov-mas.pdf.
Statement on OpenAccess to Research Publications from the National Research Foundation (NRF)-Funded Research, https://updatalibrarian.files.wordpress.com/2015/02/nrf_open_access_statement_19jan2015.pdf.
University of Cape Town, UCT Research Data Management Plan, http://www.lib.uct.ac.za/uct-research-data-management-plan.
US Code of Federal Regulations, Chapter 2, Part 215 (2006).
US Department of Energy, Office of Science, Statement on Digital Data Management, Suggested Elements for a Data Management Plan, http://science.energy.gov/funding-opportunities/digital-data-management/suggested-elements-for-a-dmp/.

9. Language Resource References

Bird, Steven, and John Bell. (2001). Grassfields Bantu Fieldwork: Ngomba Tone Paradigms, distributed via Linguistic Data Consortium, LDC2001S16, ISLRN 147-689-240-962-1.
Bird, Steven. (2003). Grassfields Bantu Fieldwork: Dschang Lexicon, distributed via Linguistic Data Consortium, LDC2003L01, ISLRN 880-081-036-797-6.
Bird, Steven. (2003). Grassfields Bantu Fieldwork: Dschang Tone Paradigms, distributed via Linguistic Data Consortium, LDC2003S02, ISLRN 973-117-906-652-9.
Brants, Thorsten, and Alex Franz. (2006). Web 1T 5-gram Version 1, distributed via Linguistic Data Consortium, LDC2006T13, ISLRN 831-344-220-094-6.
Du Bois, John W., et al. (2003). Santa Barbara Corpus of Spoken American English Part II, distributed via Linguistic Data Consortium, LDC2003S06, ISLRN 951-825-759-886-6.
DuBois, John W.,and Robert Englebretson. (2004). Santa Barbara Corpus of Spoken American English Part III, distributed via Linguistic Data Consortium, LDC2004S10, ISLRN 801-946-303-326-6.
Du Bois, John W.,and Robert Englebretson. (2005). Santa Barbara Corpus of Spoken American English Part IV, distributed via Linguistic Data Consortium, LDC2005S25, ISLRN 659-853-066-274-9.
Graff, David. (2002). The AQUAINT Corpus of English News Text, distributed via Linguistic Data Consortium, LDC2002T31, ISLRN 153-002-267-999-9.
Martell, Craig, et al. (2004). FORM1 Kinematic Gesture, distributed via Linguistic Data Consortium, LDC2004V01, ISLRN 787-443-746-101-0.
Palmer, Martha, et al. (2004). Proposition Bank I, distributed via Linguistic Data Consortium, LDC2004T14, ISLRN 874-058-423-080-1.
Reppen, Randi, Nancy Ide, and Keith Suderman. (2005). American National Corpus (ANC) Second Release, distributed via Linguistic Data Consortium, LDC2005T35, ISLRN 797-978-576-065-6.