Data Management Plans and Data Centers

Denise DiPersio, Christopher Cieri, Daniel Jaquette

Linguistic Data Consortium

- A Data Management Plan (DMP) explains how research data will be created, shared and maintained
  - Based on the notion that research data and results should be broadly accessible to the public at reasonable cost
- Many funding agencies around the world require DMPs to be included in research proposals
  - US: National Science Foundation, US Department of Energy
  - EU: Horizon 2020
  - UK: Arts and Humanities Research Council, Economic and Social Research Council
  - Australia: Australian Research Council

- Researchers must address:
  - What does the plan cover?
    - Data, metadata, software . . .
  - Where will data be deposited?
    - Designated archives, institutional website
  - Will data be archived, distributed or both?
    - Immediate v. postponed access
  - Who pays costs for archiving and distribution?
    - Funder pays, researcher pays, user pays
  - Special issues: IPR, privacy, sensitive material
    - Conditional access

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

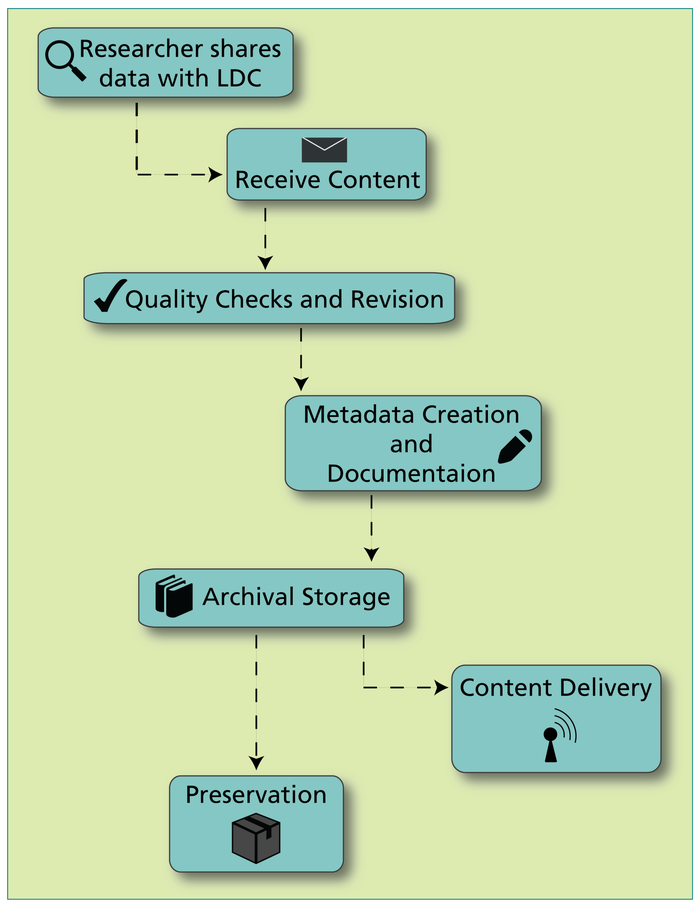
- Data Repositories: overview of archiving, curation, and distribution standards and best practices

- Data Seal of Approval (DSA)
  - Dutch Data Archiving and Networked Services
  - Repositories self-assess against various factors: formats, metadata, access, preservation, infrastructure, user policies

- Research Data Alliance (RDA)
  - International body seeking to reduce barriers to data sharing and to promote data driven innovation

- RDA Working Group: DSA-WDS (World Data System) Working Group on Repository Audit and Certification
  - Establishing common requirements for repository certification at the “core” level
  - Core certification: repository self-assessment reviewed by “community peers”
  - Incorporates existing DSA and WDS guidelines
    - Addresses organizational infrastructure, digital object management, technology
  - Replaces DSA certification in 2016

- LDC: the first and most active language resource data center
  - Growing catalog of 600+ holdings
  - Metadata based on Dublin Core standard as extended by OLAC (Open Language Archives Community)
  - Permanent repository for deposited data sets supported by state-of-the art storage and backup systems
  - Many corpora are benchmark publications and evaluation data sets used continuously by the community over more than two decades
  - LDC’s data curation process prepares corpora for distribution and preservation



- Data centers are well-positioned to administer data management plans

  - Committed to the motivating principle for DMPs: providing broad and affordable access to digital data

  - Able to exploit existing infrastructure and processes for reviewing, storing and distributing resources
    - pre-publication review
    - comprehensible data descriptions
    - improved discoverability (identifiers, metadata, sharing information across catalogs)
    - established communication outlets
    - deposit copy for benchmarking
    - regulatory expertise
    - applying best practices across the board

  - Provide assistance with cost development and budgeting: balancing funds between research needs and the goal that research data remains accessible and intact

- Open Issues
  - The DMP’s intended audience
    - How data will be prepared and presented may depend on audience training, expertise, infrastructure access
      - Different versions, formats may be needed
  - Effect on legacy data
    - Harmonizing pre- and post-DMP distribution schemes
  - Implementing cost reductions
    - Re-evaluate fixed and variable costs
    - Declining archiving costs: storage, bandwidth?
    - Cost of human services is variable
      - Licensing, regulatory matters, customer care
      - Ensuring data integrity
  - Trusted repositories not defined
    - Some standards, certifications
    - Data centers can benchmark against these
  - Persistent identifiers
    - What do they identify?
    - Not all are accepted or widespread
    - Level of granularity: the corpus, the metadata, the files