

Certifying CISER! A Data Seal of Approval Case Study

Stuart Macdonald

Cornell Institute for Economic and Social Research (CISER) /
University of Edinburgh

stuart.macdonald@ed.ac.uk



Bridging the Data Divide:
Data in the International Context

MINNEAPOLIS, MN

June 2 - 5



6-month secondment as CISER Data Services Librarian (Oct. 2013 – April 2014)

Co-ordination of the CISER Data Archive application through self-assessment for Data Seal of Approval accreditation

- CISER
- Stages and approaches
- Lessons, observations, benefits

Cornell Institute for Economic and Social Research (CISER)

Established in 1981 CISER is home to one of the **oldest** university-based social science data archives in the United States.

CISER's mission is to anticipate and support the evolving computational and data needs of Cornell social scientists and economists throughout the entire research process and data life cycle.

Collection of public and restricted-use numeric datasets to support quantitative research

- c. **27,000** online files in addition to thousands of studies on CD/DVD

Emphasis on studies that match the interests of Cornell researchers including:

- demography (state/federal censuses), economics, health, labor, election studies, attitudinal and behavioral studies
- family life etc.

Reference and consultation service to match user needs with appropriate data: finding, accessing, using data

Cornell researchers can download Winzipped data files from online catalog in formats conversant with statistical analysis software

Let's **start** at the very beginning




Gain familiarity with:

DSA Guidelines
(applicants & reviewers, statements
of compliance)

Five fundamental criteria
(online, accessible, usable, reliable,
citeable)

Three stakeholders
(producer, repository, consumer)

General information:

-  DSA information booklet (1.0 MB)
-  DSA leaflet (1.0 MB)
-  DSA overview article (905.7 KB)

DSA Regulations:

-  DSA regulations (343.3 KB)

DSA Guidelines:

-  Guidelines 2014-2015 (434.3 KB)

Identify a cross-section of **successful** DSA applications and gain familiarity with content

• ICPSR, Odum Institute, 3TU.DataCentrum, UKDA, ADS

Assemble draft spreadsheets containing statements from successful applications pertinent to CISER Data Archive

Couple with requirements from the Applicant Manual for each statement

Iteration 1

C2 This statement should give a descriptive overview of the repository including background, remit, coverage, partnerships, legal status, relationship with host organisation, collaborations, relevant U		
A	B	C
3. Data producer provides the data together with the metadata requested by the data repository	4. Implemented. This guideline has been fully implemented for the needs of our repository	Examples of public facing documents include: Adequate metadata to be captured through Data Deposit forms (should collect sufficient metadata to enable re-use by search interface) or ingest procedures manual; Evidence for use of metadata standard (DDI/SDMX/MARC); both study and file-level information where possible; provisions to contact data producer if insufficient metadata/documentation (through manual checking of metadata); Metadata policy; Existence of metadata harvesting protocols (OAI-PMH)
4. Data repository has been explicit in the area of digital archiving and promulgates it.	4. Implemented. This guideline has been fully implemented for the needs of our repository	Examples of public facing documents include: Mission statement (integration of which into policies, procedures, and practices), Annual Report, Strategic Plan, Preservation policy. Evidence of outreach activities incl. presentations, meeting of representation on professional associations, other dissemination activities, promotional materials (brochures etc); succession plans (naming other organisations/partners); social media; training
5. Data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects	4. Implemented. This guideline has been fully implemented for the needs of our repository	Express CISER's relationship in relation to Cornell University (from a legal standpoint); enunciate contractual infrastructure/legally binding agreements (between producer, repository, consumer); public statements in relation to protection of human subject and potentially sensitive/confidential information; consent forms; assertion of copyright and IP; procedures for breaches of contract; Terms of use & restricted use contracts. Data access rights (public, Cornell, restricted); software/infrastructure that will provide a secure virtual data enclave for confidential data; staff trained in handling restricted-use data; articulate how confidential data is stored, detail national/international laws under which CISER operates in relation to privacy and protection of research participants (e.g. Office for Human research Protections, University Review Boards, HIPAA etc); Articulate Data Security Policies (describing laws governing CISER's handling of data containing personally identifiable information and protected health information and the mechanisms in place to mitigate disclosure risks through rigorous data security measures (usage restrictions, policy protections and technological protections); detail Cornell Administrative policies as well as other ICT security policies; Ethics training (and certification) for staff working with sensitive data
6. Data repository applies documented processes and procedures for managing data storage	4. Implemented. This guideline has been fully implemented for the needs of our repository	Examples of public facing documents include: Preservation Policy (complete with details); Quality Control and Quality Assurance procedures across the data lifecycle; Data back-up plans/strategies (complete with details); Disaster recovery plans/procedures (tech & human); data migration/emulation processes. Highlight any third party involvement (incl. roles and responsibilities) re. storage/back-up. Evidence of ISO standards (e.g. Quality Management ISO 9000 series)
7. Data repository has a plan for long-term preservation of its digital assets	3. In progress. We are in the implementation phase	Primarily evidence required of Preservation Policy/Framework (migration, normalisation, format monitoring, versioning - as per OAS); reference to Appraisal policies (re. integrity, completeness and authenticity of data submissions); Use of Checksums. Evidence of future proofing
8. Archiving takes place according to explicit work flows across the data lifecycle	3. In progress. We are in the implementation phase	Existence of policies and procedures that follow the archival lifecycle complete with predetermined criteria that apply at each stage (e.g. ingest or deposit/ collection development policies, appraisal procedures, metadata/catalog policies, storage/back-up plans etc); staff training and expertise
9. Data repository assumes responsibility from the data producers for access and availability of the digital objects	4. Implemented. This guideline has been fully implemented for the needs of our repository	Licence agreement between producer and repository at ingest stage (granting archive with permission to distribute, catalogue, store, copy, reformat, preserve and disseminate data collections) - this may be via a Data Deposit Agreement included in Deposit form. Articulate Restricted-use agreements also (for confidential data access) & role of producer as proprietor/joint-proprietor (e.g. whether producer distribute data independently). Risk management and disaster recovery procedures also
10. Data repository enables the users to discover and use the data and refer to them in a persistent way	3. In progress. We are in the implementation phase	End User Agreements & Access conditions (may differ according to data collection - red/yellow/green traffic light); Data catalog through search interface (with standard features) allows access to metadata about data collections and data where applicable. Data exist in domain specific formats (SPSS, Stata, ASCII etc) that are deemed useful to consumers. Data at study level are (will be) assigned a local unique identifier (DITitle). Investigation into implementing study-level persistence using DataCite DOIs underway with EZID (give URL). Availability of analysis software for data use. CISER planning to enhance discovery metadata. Study-level citation explicit for search results

Iteration 2

Quality Guidelines	Min. required statement of compliance	CISER self-assessment statement requirements:	Public facing URLs
0. Repository Content		<ul style="list-style-type: none">This statement should give a descriptive overview of the repository including background, remit, coverage, partnerships, legal status, relationship with host organisation, collaborations, relevant URLs etc	CISER: http://ciser.cornell.edu/ CISER Data Archive: http://ciser.cornell.edu/ASPs/sea_rch.asp CISER System Usage Policies: http://ciser.cornell.edu/computing/manual/policies.shtml
1. Data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, the compliance with disciplinary and ethical norms	3 – In progress	<ul style="list-style-type: none">Data deposit formData preparation and archiving guidanceData creation and management guidesData appraisal policyMetadata & documentation policyData collection policyComprehensive advice on all aspects of the data lifecyclePublic facing URLs	
2. Data producer provides the data in formats recommended by the data repository	3 – In progress	<ul style="list-style-type: none">List of accepted and durable file formatsEvidence of quality control / validity checkingUse of format diagnostic tools (JHOVE, DROID, PRONOM)Normalisation proceduresAppraisal policyProcedure to deal with non-standard formatsPublic facing URLs	
3. Data producer provides the data together with the metadata requested by	4 - Implemented	<ul style="list-style-type: none">Data Deposit form (to capture metadata)Ingest procedures manual	CISER Computing Training: http://ciser.cornell.edu/beta/wor

Weekly meetings with CISER staff (1 – 2 hours)

Statements were assigned to members of staff with particular expertise (storage, security, formatting, restricted data, catalogue, metadata)

Separate meetings held to discuss individual assignments

Separate meetings held to update policies and craft new policies where they didn't already exist

In total 12 person **weeks** (principally my time plus colleagues)

Submission March 2014, Award July 2014

Identified ‘quick wins’ e.g. existing policies, agreements, terms of use, guideline 0

Knowledge, workflows and procedures existed:

- In people’s **heads**
- Technical documentation
- Legacy printed material (incl. policies)
- Internal and external online links

However information gathering and evaluation was an iterative process

Easy to underestimate time required to assemble and craft new policies (such as Preservation and storage, Security, Versioning, Data Collection), mission statement , and other public facing documentation

Proofreading, consistency of language, terminology, and narrative was also time consuming (‘different voices’)

Organisational and community **benefits**

Entire exercise invaluable for clarifying and articulating organisation's archival practices

Promoting trust and confidence between the three stakeholders in the data supply chain

- all are working to a common set of standards or principles

Easier to conduct systematic review of technical / human processes and procedures in future

As/when new compliant tools, technologies, standards emerge the archive will be better equipped to respond to necessary changes in data stewardship workflows

The application process helped to identify service gaps and areas for improvement/modernisation in archival process and procedure

Raise the profile of the archive and preservation with Cornell senior managers

Useful for new archive staff to obtain holistic perspective on the mechanics of a mature data archive

Sound foundation for further institutional **TDR** exercises such as DIN 31644 (34 metrics) & ISO 16363 certificate or TDR Checklist: c. 100 metrics

Highlights areas of interworking and interaction between archival colleagues for purposes of streamlining of operations

Contributes to the social science data archival community and the data stewardship profession by openly sharing archival processes and procedures

In **summary** the process of applying for Data Seal of Approval was a positive learning experience for archival staff and the organisation as a whole.

But more importantly it is a public pronouncement of archival intent:

to demonstrate reliable and trusted access to managed research data for its academic community, both now and into the future.

Thank You.

Acknowledgements:

Bill Block

Warren Brown

Florio Arguillas

Jeremy Williams

Janet Heslop

Irene Hawes

Lynn Martin

Ben Perry