CASE STUDY

A Digital Curation Centre Case Study
August 2015

Reviewing Research Data Platform Capabilities at Cornell Institute for Social and Economic Research (CISER)

Florio Arguillas, Janet Heslop (CISER) and Angus Whyte (DCC)

Introduction

This case study describes repository development priorities identified by the Cornell Institute for Social and Economic Research (CISER).([1]) Founded 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 social scientists and economists, based at Cornell University and elsewhere, throughout the entire research process and data lifecycle.([2])

The case study is one of three accompanying the DCC guide *How to Evaluate Research Data Repository and Catalogue Platforms*. Each portrays issues affecting platform choices and implementation decisions in different institutional contexts. This case study describes how CISER used 'Data ReCap' a capability model for data repositories. Data ReCap, introduced in the DCC Guide, is intended to help research data management professionals highlight organisational (non-functional) or technical (functional) gaps in their data repository or catalogue capabilities. CISER staff used Data ReCap to support their planning of service improvements. CISER is currently reviewing its technical infrastructure, to map and translate existing tasks and functions to the OAIS reference model. Applying the Data ReCap model identified priority areas for this review.

Background and context

The CISER Data Archive houses an extensive collection of public and restricted numeric data files in the social sciences with particular emphasis on studies that match the interests of Cornell researchers: demography, economics and labour, political and social behaviour, family life, and health. Data archive functions include making data available to the broadest audience permissible; providing a secure, safe research computing environment to facilitate data access and use for researchers; and data consulting support from staff experienced in using social science data, in order to maximize the benefits of the data archive and research computing facilities. This including significant depth of available expertise in restricted data access management.

Key issues:

- Data publishing and preservation mandates
- Collection policy
- Discovery system integration
- Domain & context metadata
- Data mining & visualization
- Open interfaces
- Domain & content scope
- Discoverability
- Version control
- Persistent identification
- Preservation metadata

[1] *CISER Home page, available at: ciser.cornell.edu*
[2] *Mission statement, available at: ciser.cornell.edu/pub/policies/CISER_Mission.pdf*

In 2014, the CISER Data Archive applied for and was awarded the Data Seal of Approval.[3] The process requires a self-assessment of the archive capabilities, policies and functions as well as its responsibilities to data depositors and data consumers. Since the self-assessment is peer-reviewed, getting the Data Seal of Approval certifies that the CISER Data Archive policies and practices are sound and sustainable and that it is a trustworthy digital repository.

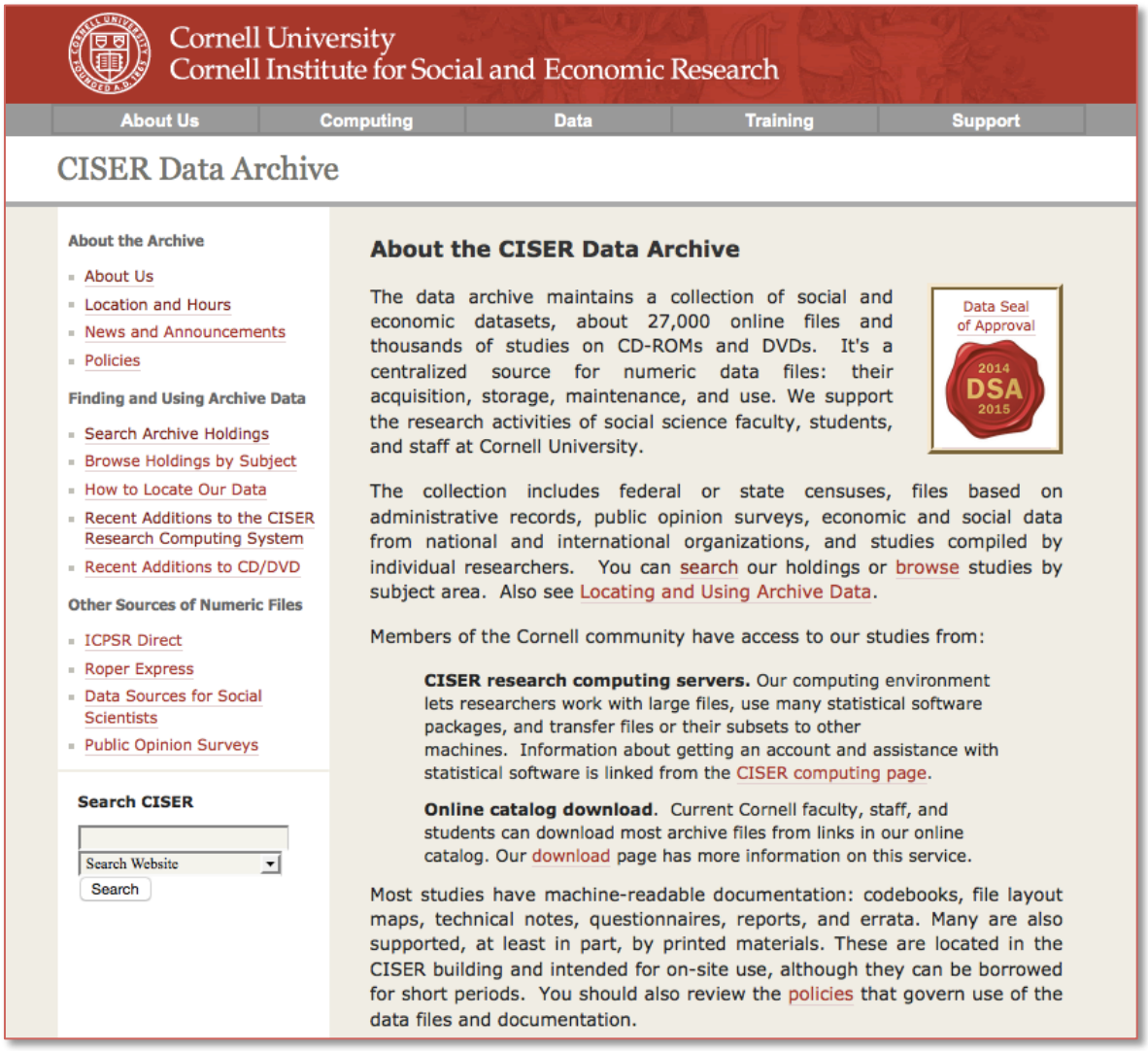

Figure 1. About the CISER Data Archive (source: www.ciser.cornell.edu/info/about.shtml)

Applying the Data ReCap model

The Data ReCap model is described in more detail in the guide *How to Evaluate Research Data Repository and Catalogue Platforms.* It is intended to help research data management professionals highlight gaps in service provision and make the case for improvements. It does this by offering brief descriptions of capabilities needed to comply with funder mandates, or meet good practice recommendations and standards. The model describes

- Organisational capabilities, considering the policy and technical environment for the data service, and drivers of stakeholder expectations.
- Data management platform capabilities, considering technical developments and user expectations.

The model consists of tables listing three levels of provision for each of the capabilities.

- Level 1 offering basic provision, consistent with funder mandates.
- Level 2 aligning with widely adopted standards, sector norms, and good practice recommendations.
- Level 3 exceeding sector norms, informed by higher levels of well-established guidelines and emerging international standards.

[3] *Data Seal of Approval, available at: datasealofapproval.org/en/*

CISER used the model to consider the organisational criteria relevant to its Data Archive service, the levels currently met, and their expectations of changes needed. Then they chose the data management capabilities important to meeting those expectations, either currently or through policy or technical development. The results are summarised in the two tables below, which show in bold the capabilities they identified as needing improved.

		Needs unmet and required level		
		Level 1	Level 2	Level 3
Needs met and current level	Not app.			
	Level 1			• **Data publishing mandate**
	Level 2		• Metadata collection policy	• **Data collection policy** • **Preservation mandate** • **Discovery system integration**
	Level 3			• Copyright & license policy • Service sharing • Outsourcing • Research intensity • Access management environment • Storage management environment • Development capability & capacity • Curation skills & capacity • Funding availability

Table 1. Research, technology and organisational environment

		Needs unmet and required level		
		Level 1	Level 2	Level 3
Needs met and current level	Not app.		• **Support for domain-specific workflows**	• **Domain & context metadata mgmt** • **Text/data mining & visualization** • **Open interfaces**
	Level 1	• Local installation • Info. security policy control	• **Access & citation tracking** • **User community Support**	• **Domain and content scope** • **Discoverability** • **Version control** • **Persistent identification** • **Preservation metadata**
	Level 2		• Access management • Service compliance and certification	• **Data collection policy** • **Preservation mandate** • **Discovery system integration**
	Level 3			• User Licensing flexibility • File format checking • Data integrity/fixity checking • Continuity support

Table 2. Data management platform capabilities.

Below we look in more detail at how CISER describes itself in terms of the model, firstly looking at the organisational expectations. After that we turn to the functional and technical capabilities currently provided, and improvements sought to fulfil the Archive's aspirations.

The research, technology and organisational environment

CISER is the social science research support arm of Cornell, a highly research intensive institution, and thus its primary activity is research support. As such, almost all the organisational characteristics listed in the model matter to CISER's service provision. The one exception is that there is no current requirement to integrate the Data Archive with an institutional CRIS (Current Research Information System).

CISER helps faculty, staff, and students engage in substantive social science or economic research to find or acquire data, and to preserve, publish, and make web accessible the data they create (see Online Catalog[4]). Cornell's Research Data Management Service Group (RDMSG) also offers a collaborative, campus-wide organization that assists with creating and implementing data management plans, applying best practices for managing data, and finding data management services at any stage of the research process.[5]

[4] *CISER Data Archive: Online Catalog, available at: ciser.cornell.edu/ASPs/search.asp*
[5] *Research Data Management Service Group, available at: data.research.cornell.edu*

Operating policies and strategies

Service compliance and certification

In 2014, CISER achieved Data Seal of Approval certification. This is based upon peer review and provides international recognition according to Trusted Data Repository (TDR) standards. This is a significant achievement and, rather than invest effort in seeking higher levels of TDR compliance CISER's priority at time of writing was to focus on service improvements that align with its client needs such as those highlighted below.

Copyright and license policy

The Archive offers a variety of license choices or custom-made terms and conditions based on data provider requirements. All CISER Data Archive users must agree to a Terms of Use Policy ([6]) prior to gaining access to data held in the archive. The policy's main conditions are that: users are responsible for complying with all applicable federal, state and local laws; they must abide by Cornell University policies; and agree to adhere to data provider licensing requirements.

Other legal contracts and regulations that CISER handles are primarily related to Data Provider Agreements (DPA) for restricted-use datasets within the Cornell Restricted Access Data Center (CRADC).([7]) The DPA stipulates use, dissemination, and backup specifications of restricted-use data. The Office of Sponsored Programs (OSP) and the Institutional Research Board (IRB) evaluate all DPAs for terms and regulations governing the protection of human subjects. ([8],[9]) In addition, the DPA includes information on penalties for noncompliance. OSP negotiates these agreements if necessary, and signs the DPA on behalf of Cornell University.

Service sharing strategy

CISER operates with significant external partnerships, including formal relationships with the Inter-university Consortium for Political and Social Research (ICPSR), and the Data Documentation Initiative Alliance. It also became a member of Data Pass in 2014. Its staff play active, if not leadership, roles in organisations such as IASSIST, Association of Public Data Users (APDU), and New York State Data Center (NYSDC) to name a few, and is responsive to assistance sought by the membership of these organizations.

Outsourcing strategy

CISER is staffed with data librarian, software, and web developers and thus all service and infrastructure functions are delivered in house. However, this does not preclude CISER from utilising tools that have been developed outside to enhance service delivery, such as StatTransfer for data conversion and Sledgehammer for metadata creation.

Data collection policy

The CISER Data Archive's data collection policy mainly applies to locally held data and metadata. Metadata scope includes externally held datasets in the archive's catalogue. The data collection policy scope aims to serve the specialist needs of external depositors, as well as those within Cornell's research community. CISER's main emphasis is on demography, economics and labour, political and social behaviour, family life, and health. The collection includes federal or state censuses, files based on administrative records, public opinion surveys, economic and social data from national and international organisations, and studies compiled by Cornell researchers.

Collected data predominantly deals with the United States, European countries and international data produced by intergovernmental organisations. Emphasis is also placed on data related to New York State, Cornell University's home state, but the Archive acquires or accepts data relating to any geographic area. Although it is often difficult to obtain data from some developed nations and most developing states, the CISER Data Archive works with researchers to meet their needs.

CISER staff appraise data based on faculty recommendations. Criteria include the quality of the data and the reliability of the distributor, and expected future use by a broad constituency of social science users. The Data Archive will attempt to acquire any set of data requested by faculty members and/or their students who are engaged

[6] *Terms of Use Policy, available at: ciser.cornell.edu/pub/policies/CISER_Terms_of_Use.pdf*
[7] *CISER Policies, available at: ciser.cornell.edu/pub/policies/CISER_Policies.shtm*
[8] *Office of Sponsored Programs, available at http://www.osp.cornell.edu/*
[9] *Institutional Research Board, available at http://www.irb.cornwll.edu/*

in substantive social science or economic research, within organisational policies regarding cost, quality, restrictions, and expected future use by a broad constituency of social science and economics users.

While data acquisition is primarily demand driven, the Archive is also pro-active, aiming to anticipate demand. CISER Research Staff also create datasets that would have high demand or potential to save researchers time and money. Download centres created for the U.S. Census 2010 Summary Files are an example of that.([10]) The Archive also maintains externally published data series identified as core to the data collection. Due to contractual agreements between Cornell University and the Inter-university Consortium for Political and Social Research (ICPSR) members of the Cornell Community are entitled to obtain any of the data offerings of the Consortium. CISER Data Archive actually serves all members of the community in terms of data acquisitions from the Consortium, regardless of subject area.

Metadata Collection Policy

Where possible data are accompanied by comprehensive machine-readable documentation: codebooks, file layout maps, technical notes, questionnaires, reports, and errata in open and accessible formats. In cases where documentation is incomplete, the Archive staff work with data producers to gather more, to ensure that data files are useable and understandable. Hardcopies are converted into electronic form using the PDF/A format, and made available if machine-readable documentation is not. The Data Archive reserves the right to reject datasets deemed inadequately documented.

Metadata creation continues across the data lifecycle (i.e., from data conceptualisation to collection, processing, distribution, discovery, analysis, repurposing, and archiving). If necessary, additional user information is provided, such as a Readme file or other documents that detail the changes that were made to the original data and/or other instructions for using the collection.

The Archive will continue to broaden its metadata collection policy by collecting and making discoverable metadata at the variable level.

Preservation mandate

The Archive's preservation mandate is very important to CISER's mission. Copies of the collection (e.g., data and metadata) are held locally, backed up to a separate subnet and preserved via an offsite backup. Currently the service provides metadata and file backup, and commits to file format checking at ingest and to fixity or data integrity. The Archive also performs some format migration and wants to grow its service to migrate further formats and keep content readable and compatible into the future.

Automated scripts are run regularly for verification. This includes auditing file permissions; reporting on all file changes since the previous validation; and checking the archive path and filename identified in the catalog metadata. Data versioning criteria are consistently applied to changes in files and documentation (such as error correction, additional variables, changed access conditions, format changes). Once deposited, files in datasets are never changed and only minor changes to the metadata are allowed. This often involves working closely with the data producer. Changes to the data themselves, or major changes to metadata, are issued as a new version of the dataset. Changed datasets will be issued a new study-level persistent identifier (DOI) using the California Digital Library's EZID service.([11])

Data publishing mandate

CISER mandate for data publishing is evolving. Although currently limited to making data or metadata web accessible, publishing effort will grow as the Archive begins to implement the DDI (Data Documentation Initiative) standards for data citation and metadata exchange. This will make data and metadata more discoverable and machine-readable.

All the Archive's data studies are assigned a locally generated unique identifier, and will be assigned a study-level DOI using the EZID service (except those prohibited by the data provider in the case of restricted access files). Documentation provided with each study includes a standard format study-level citation.

---

[10] *E.g. Census of Population, 2010: Summary File 1 (SF1) Download Centre, available at: ciser.cornell.edu/ASPs/search_athena.asp?IDTITLE=2577*
[11] *California Digital Library EZID, homepage available at: ezid.cdlib.org*

The technology environment

Access management

Access is based upon a "green-yellow-red" light system. The files that are publically available have a "green light", those identified with a "yellow light" are limited to Cornell affiliated researchers only, while those classified with a "red light" are restricted, and access to these is permitted according to the terms stipulated by the data providers.

The Archive is locally hosted, and accessible either by searching and browsing the data catalogue, or via a CISER Research computing account. The user will be able to identify which datasets they are able to access via the "green-yellow-red" light system. Anyone wishing to download a dataset must accept a Terms and Conditions statement. Access via the Research computing system (CISERRSCH) allows the researcher to work with the archive data files and perform analysis, for which many preinstalled statistical software packages are available. The data files are accessible via file-level permissions, as mandated by the data providers.

For public-use datasets, CISER complies on a case-by case basis with data producer terms and conditions through data producer agreements signed by the Data Librarian. CISER also has annual memberships with other data producers including ICPSR ([12]) and the Roper Center for Public Opinion Research ([13]), allowing it to make their datasets available.

Discovery systems integration

The online catalogue offers robust search and browse facilities to enable discovery of and access to both public-use and restricted-use data files, including legacy data held on CDROM/DVD. Users are also able to download codebooks and other documentation materials through the catalogue. Users can search for data by title, producer, and principal investigator in addition to conducting free text searching with truncation. The holdings are also browseable by subject area. Some of the collection is searchable via the University Library catalogue, and the goal is to extend this to all of our holdings so they are identifiable via study-level DOIs.

Storage management environment

Preservation is dependent on CISER's storage infrastructure, which is therefore designed to accommodate scalability, reliability, and sustainability, and to meet quality control specifications and security regulations. The Data Archive is stored on network-attached storage (NAS) in both compressed and uncompressed format in Cornell University's Data Centre. The compressed data is for public download access via the CISER data catalogue. The uncompressed data files, documentation, and ancillary files are housed on research-computing servers, which allow CISER account holders to locally prepare, analyze and manage data using statistical software packages.

Resourcing and funding aspects

Development capability and capacity

This is extensive, and includes a full-time Data Librarian as core member of staff. This role is committed to spearheading the development and enhancement of CISER's data-rich environment, including the collection of social and economic datasets, acquisition and ingest of new datasets, and the design and implementation of the Archive's structural improvements/ innovations of the archive. In addition, the Data Librarian also leads accreditation activities (i.e. Data Seal of Approval, Data-PASS), promoting and adopting social science metadata standards, and continued development and evolution of Data Archive policies (i.e. data citation, documentation, and metadata polices).

Curation skills and capacity

The Archive staff have broad-based skills, although capacity is one of the challenges discussed below. CISER's full-time Data Librarian has an all-encompassing background in data management and curation. In addition, two other staff members assist with Data Archive projects; an Application Programmer, who supports data processing and cataloguing functions, and a Systems Engineer who helps develop tools to incorporate data ingest, DOIs, and innovation projects.

[12] *Inter-university Consortium for Political and Social Research (ICPSR), homepage available at: www.icpsr.omich.edu/icpsrweb/landing.jsp*
[13] *Roper Center for Public Opinion Research, homepage available at: www.ropercenter.uconn.edu/*

Funding availability

All three staff roles (Data Librarian, Applications Programmer, and Systems Engineer) are fully funded in CISER's annual budget.

Data management platform capabilities and current challenges

This section describes how the CISER Data Archive currently delivers capabilities selected from the model, focusing on those areas highlighted in Table 2 as ones where improvements are sought.

Policy control and reporting

Broadly CISER is satisfied with its capabilities in this area.

Information security

Authentication is not required for access to public-use datasets, if accessing via the CISER web catalogue. However, a Terms and Conditions must be accepted in order to download the data files. Where the data provider obligates, the user would be required to authenticate with CUWebAuth (Cornell Netid required) via the web catalogue or a CISER computing account.

Access to the data files utilizing a computing account is strictly controlled via Microsoft Windows New Technology File System (NTFS) through individual level file and folder permissions as outlined in the CISER Data Security Policy. Although CISER has the potential to log every file access, edit, and/or deletion, it has chosen not to do so due to the extensive storage requirements of the log files.

Content organization, publishing and preservation

Domain and content scope

This scope is growing, however it is currently limited to single-file datasets, with generic metadata and minimal support for specific content types. The ultimate goal is to broaden the scope to not only support multiple-file datasets, but also configurable metadata, diverse content types and sizes.

Support for domain-specific workflows

The Archive would like to provide configurable workflows for ingest and dissemination that we can control for domains, user groups, or object types, but we do not currently have this capability. CISER staff who manage data have a set of internal guidelines that they adhere to for ingest as well as dissemination of metadata and collection. Other processes such as long-term preservation (e.g. normalization, version control, sustainability) are detailed in the CISER Data Preservation and Storage Policy.([14])

Discoverability

Discoverability matters to the CISER Data Archive and we aim to improve it in our platform to also provide support for linked open data standards.

Domain and context metadata management

The Archive is currently developing more extensive capabilities to support editing of domain-specific metadata or associate controlled vocabulary terms with data/metadata records.

Version control

Version control matters to the Archive, and work is underway to improve the platform's capabilities, so change history can be recorded and made visible, and multiple authorized users will be able to edit items at the same time and record changes. The Data Archive platform version control capability is currently limited to time-stamping dataset uploads and recording latest edits. However, the Archive has a set of guidelines for distinguishing between significant and minor changes to a dataset and documentation to determine whether a new version of the dataset and its corresponding documentation is merited. These guidelines ensure that data versioning criteria are consistently applied to

[14] *CISER Policies, available at: ciser.cornell.edu/pub/policies/CISER_Policies*

changes in data files and data documentation (including correction for error, amendments, additional variables and/or records, changes in access conditions, format changes ([15]).

Persistent identification

The Archive is (at time of writing) in the process of adopting study-level DOI using the EZID service, except for those where access is prohibited by the data provider.

Preservation metadata

The Archive wants to extend its capability to support automated extraction of technical or descriptive metadata to standard schema or vocabularies. Metadata is already stored separately from file backup location.
File format checking, data integrity/fixity checking, and continuity support

File format checking, data integrity/fixity checking, and continuity support

File format checking, data integrity/fixity checking, and continuity support are important to the Archive and its current platform meets requirements. It supports file format validation at ingest, and file normalization to selected preservation formats. File formats can be migrated to new versions, or new formats when these become widely available. The platform generates checksums using an MD5 File Hasher utility for every data file added to its collection to ensure the integrity of the digital file both now and into the future. MD5 checksum validations are executed to report all files that have been added, deleted, or modified since the previous validation**Integration and interoperability**

Open interface

An ingest tool for custom data ingest and access to workflows, machine-readable metadata, linked open vocabularies, etc., is being planned for future implementation.

User community support

User community support is an area that CISER intends to improve, and we anticipate that an advisory group will be established within the near future to provide input to our development roadmap. In addition, our FAQs could use some updating and outreach efforts need to be expanded within our community.

Lessons learned and challenges for the future

The CISER Data Archive continues to improve, develop, and align its services as it endeavours to achieve its mission of anticipating the evolving needs of researchers on campus. Currently CISER is in the process of reviewing its technical infrastructure, to map and translate existing tasks and functions to the OAIS reference model. In the short term, priority items for the development roadmap include a data ingest tool with improved usability, and more effective checking of study-level metadata completeness.

Upgrading an established data archive can be challenging in terms of staff resources. Staffing challenges have delayed development. The Data Librarian position was unfilled for a long of time, stretching staff capacity and capabilities. This had cascading effects on the day-to-day work of the Archive.

Meeting international standards can be very demanding, and archives seeking to improve their service may need to prioritise which standards and elements of those they comply with based on evidence of user and organisation needs. CISER found the DCC Data ReCap model very helpful to identifying gaps between capabilities met and unmet, to inform its development roadmap.

Please cite as: Arguillas, F., Heslop, J. and Whyte, A. (2015). 'Reviewing research data platform capabilities at Cornell Institute for Social and Economic Research (CISER)'. *DCC Publications*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/case-studies

[15]*CISER (2014) Data Archive Versioning Policy, available at: http://www.ciser.cornell.edu/pub/policies/CISER_Policicies.shtm*