ARCHIVES AND MANUSCRIPTS, 2016
VOL. 44, NO. 2, 73–85
http://dx.doi.org/10.1080/01576895.2016.1179585

Routledge
Taylor & Francis Group

Organisation and description of datasets

Jinfang Niu

School of Information, University of South Florida, USA

ABSTRACT
Through the exploration of the websites and catalogues of a number of national archives and dataset repositories, this article identifies the similarities and differences between the knowledge organisation methods of national archives and those of dataset repositories, and finds that dataset repositories use more specialised metadata standards and support more flexible knowledge organisation and easier data access and use. Based on the findings, the author discusses some possible approaches that archival institutions can take to improve data description and support better data discovery and use.

KEYWORDS
Data curation; dataset; digital archiving; metadata

Introduction

In the past few decades, the widespread use of relational databases has converted large quantities of paper records into structured datasets. In addition, the data-sharing policies of governments and journals have urged researchers to share and deposit data to libraries and archives for long-term preservation. It is inevitable that archival institutions will have to deal with more structured datasets. However, the archives community does not seem well versed in the organisation and description of datasets. Archival description usually focuses on describing various levels of aggregates, such as record groups, series and file units. For the item level, archival description standards, such as General International Standard Archival Description (ISAD(G)), Describing Archives: A Content Standard (DACS) and Encoded Archival Description (EAD), define general elements that apply to all kinds of records, such as title and date. Archival description standards pay little attention to the special features of records in particular formats and genres, such as the playing speed of sound recordings and the polarity and colour of photographs. When the special features of certain types of records need to be described, specialised metadata standards are recommended. For example, DACS recommended Resource Description and Access (RDA) for publications and various other specialised standards for moving images, sound recordings, rare books and manuscripts.[1] In the 2013 version of DACS, the recommended standard for describing datasets is 'Federal Geographic Data Committee. FCDC-STD-001-1998. *Content Standard for Digital Geospatial Metadata* (revised June 1998). Washington, DC: Federal Geographic Data Committee, 1998. http://www.fgdc.gov/metadata/csdgm'.[2] In fact, this standard will be replaced by ISO 19115 and its accompanying standards.[3] In addition, this is a standard only

CONTACT Jinfang Niu Jinfang@usf.edu

for geographic datasets. There are many other metadata standards for describing datasets in general, datasets in social sciences and datasets in various scientific disciplines.

About 15 years ago, Shepherd and Smith (2000) examined the suitability of ISAD(G) for the description of datasets.[4] Based on their investigation of dataset descriptions at the National Digital Archive of Datasets (NDAD) and a number of other data archives, they adjusted and extended ISAD(G) and produced guidelines for archivists to catalogue electronic datasets. The guidelines constitute a special-format cataloguing manual for electronic datasets. This is the only known specialised description standard for datasets created based on an archival description standard. However, NDAD was discontinued in 2010,[5] and its dataset description guidelines are not used by the UK National Archives today.[6]

Compared with archival institutions that manage datasets as one of their many kinds of resources, dataset repositories are specialised in managing and preserving datasets. Thus, dataset repositories are likely to have more advanced or specialised methods and tools for dataset management. In light of this, this study will try to find out how national archives and dataset repositories organise and describe datasets and support dataset discovery and use, how the knowledge organisation methods and information discovery tools of national archives are similar and/or different from those used by dataset repositories, and then decide what archival institutions can do to improve their dataset management practices. Findings from this study will inform archivists and help them deal with the challenges of data curation.

What are datasets?

NDAD defines datasets as 'collections of raw data or information which have been removed from their original computing environment (databases) and can naturally be represented as "a series of tables containing columns for particular types of information and rows for each instance of data"'.[7] This definition differentiates datasets from databases: 'Databases generally exist as working systems in themselves (containing complicated internal relationships) and are often "active" in nature (the boundaries of the database and the data within it are continually changing)',[8] whereas datasets are exported from databases. While theoretically this is a sound distinction between datasets and databases, in practice the term 'database' is often used for data exported from the live environment and preserved by archives or data repositories. In fact, many databases were found using the catalogues of national archives examined in this study.

The NDAD definition limits the scope of datasets to tabular data. However, many data repositories store other kinds of structured data, such as linked data that consists of Resource Description Framework (RDF) triples, as well as network data that consists of edges and vertices. For example, many datasets encoded in RDF/XML can be found in data.gov and there are many network datasets at the University of California-Irvine Network Data Repository (https://networkdata.ics.uci.edu/index.php) and the Stanford Large Network Dataset Collection (http://snap.stanford.edu/data/).

The definition from the Long Term Ecological Research Network is broader: a dataset means 'Digital data and its metadata derived from any research activity such as field observations, collections, laboratory analysis, experiments, or the post-processing of existing data and identified by a unique identifier issued by a recognized cataloging authority such as a site, university, agency, or other organization.'[9] In this definition, datasets are defined based

on their provenances in research activities, rather than based on their structure or source from relational databases. This definition still limits the scope of datasets to digital format.

In the broadest sense, the term 'dataset' is used to refer to a collection of any kind of resources used for analysis and research purposes, whether or not in digital format, structured or not, numeric or not. A collection of specimens, images, audio and videos or paper records can be called a dataset as long as they are analysed and used to draw conclusions for a research project. For example, the UK Government Web Archive is considered a large dataset for researchers who analyse it and draw conclusions from it via data analysis.[10]

Although there are many kinds of datasets and the scope of datasets can be very broad, tabular data is the most common type of data preserved by the institutions investigated in this study. A dataset often includes a number of data files and documentation files. Documentation files are often user manuals, data dictionaries, coding schemes and other items that help users understand and use data. They usually contain more detailed information about datasets than metadata recorded in catalogues.

Methodology

This study was conducted through examining the websites and catalogues of four national archives and 12 dataset repositories including four social science data archives, four scientific data centres and four newly emerged government open data portals. National archives were chosen because they usually have more resources and expertise than other types of archival institutions. Thus, they are likely to represent the best practices in managing datasets among archival institutions. This study selected the national archives of four English-speaking countries in the developed world, including the National Archives and Records Administration (NARA) of the United States, the National Archives (TNA) of the United Kingdom, Library and Archives Canada (LAC) and National Archives of Australia (NAA). National archives preserve many types of records. To study the description of datasets, the catalogues of these institutions were searched using the terms 'dataset' and 'database'. Catalogue records and documentation of data files, if available, were analysed. Other parts of the websites of these national archives were also explored to identify relevant information.

The four social science data archives include the Interuniversity Consortium for Political and Social Research (ICPSR), the UK data archive, the Longitudinal Internet Studies for the Social sciences (LISS) panel data archive (http://www.lissdata.nl/dataarchive/study_units) and the Harvard Dataverse Network (https://thedata.harvard.edu/dvn/). The first two are very large and established data archives with a long history. The Harvard Dataverse Network is large, but relatively new. It is very large because it integrates data from existing data archives, such as ICPSR and the Murray Research Archive. It is new because the earliest dataverse was released in 2007. Although primarily a social science data archive, it also contains some scientific data, such as astronomical data.[11] Its webpage states that it is 'open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data.'[12] The LISS panel data archive is a small and relatively new data archive. It preserves and releases data gathered through the MESS project (Measurement and Experimentation in the Social Sciences), which conducts two panel studies, the LISS panel (started in 2007) and the Immigrant panel (started in 2010).

The four scientific data repositories are the Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/), the Long Term Ecological Research Network (LTER) (https://portal.

lternet.edu/nis/home.jsp) data portal, the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) (http://daac.ornl.gov/) and Dryad (http://datadryad.org/). The first three have a relatively longer history and provide access to data mainly from government-funded research centres. The LTER Network was created by the National Science Foundation of the United States in 1980. This network consists of a number of research sites across the United States, each of which maintains its own data repository. The LTER data portal integrates data across all of these sites. Similarly, GCMD is also an integrated data catalogue that does not store data locally, instead it links to data and documentation files stored externally in other data repositories. ORNL DAAC for biogeochemical dynamics is one of the NASA Earth Observing System Data and Information System (EOSDIS) data centres.[13] Dryad is relatively new and allows individual researchers to upload and share their data. It preserves research data underlying scientific and medical publications.

The four newly emerged government open data portals are from the same countries as the national archives. They are the government open data portals of the US (data.gov), UK (data.gov.uk), Canada (open.canada.ca/en/open-data) and Australia (data.gov.au). The four government open portals are primarily catalogues that point to external data stored in the repositories of government agencies. Government open data portals are related with national archives in certain ways. Datasets released through government open data portals are government records, hence under the control of records schedules approved by the national archives.[14] Some of the datasets with archival value may be transferred into the national archives for preservation. On the other hand, national archives, as government agencies, may release their own data through government open data portals. For example, as of November 2014, NARA has released seven datasets through data.gov, such as the XML raw data files exported from its online catalogue. Similarly, LAC has released five datasets through open.canada.ca/en/open-data, such as the comma-separated value (CSV) file exported from the database 'Faces of the Second World War'. It is noteworthy that these datasets are created by the national archives, not produced by other government agencies and then preserved by the national archives. For example, the database 'Photographs: Canadian Nurses' of the LAC was created by digitising old photographs from a number of archival collections. It contains item-level description for each photo and is searchable by keywords and themes. Releasing raw data in open standard format through government open data portals enables other ways to use the data in addition to searching for individual records within the database. For example, they may be analysed statistically or converted into RDF/XML format and integrated with other linked data sources.

The 12 dataset repositories, although a small sample, are diverse enough to cover major types of dataset repositories available. The catalogues of these dataset repositories were examined to identify the metadata schemas used, formats of data and documentation files, and data analysis tools provided. The websites of these dataset repositories were explored to identify other means to support dataset discovery and use. Where necessary, external websites were also used to gather information about the metadata schemas used by a dataset repository. For example, the metadata schemas used by two of the government data portals are based on the Data Catalog Vocabulary (DCAT). The W3C website was explored for more details about this vocabulary. After all the relevant information was gathered from all the institutions, the knowledge organisation methods and information discovery tools used by national archives and dataset repositories were compared and the similarities and differences were identified.

Similarities between national archives and dataset repositories

Both dataset repositories and national archives use metadata-based catalogues and browsing structures in order to support dataset discovery. In addition, they both support multi-level description, authority control and entity linking in their catalogues.

Multi-level description

Like many other archival institutions, the catalogues of the four national archives support multi-level description. For example, at NARA, the description of databases, their context and components (data files and documentation files) are organised into a multi-level structure and the user is able to navigate up and down the hierarchical structure following hyperlinks. Multi-level description is also commonly found in dataset repositories. The four government open portals all use the open source software Comprehensive Knowledge Archive Network (CKAN) (http://ckan.org/features/) to power their data catalogues. CKAN supports a two-level description. The primary unit of cataloguing is a dataset. Users can also describe each item within a dataset. The items usually include data files, metadata and documentation files, and various other relevant items. For example, the dataset 'Consumer Complaint Database' found in data.gov includes one csv file, one JSON file and one XML file. Metadata elements for the dataset level include title, publisher, contact, access and use information, number of views and so on. Metadata elements for the item level include URL, source, number of views, rating, format, creation date, data of last update, licence and so on. Similarly, the metadata application profile of the Dryad repository describes data packages, as well as data and documentation files in the data packages.[15] The four social science data archives primarily catalogue datasets at the study level. In the social science data documentation standard Data Documentation Initiative (DDI), a study is defined as 'a single coordinated set of data collection/capture activities, such as a one-time survey or a single iteration of a multi-year repeated study (such as one year of a longitudinal survey)'.[16] In the case that a study belongs to a series, these data archives usually create a separate metadata record for the series, which is linked to the metadata record for each individual study. One example of a series at the UK data archive is the 'English Housing Survey', which consists of many studies. Multi-level description is also used by scientific data repositories. The dataset catalogue of GCMD uses the metadata standard Directory Interchange Format (DIF) to describe datasets. The <Parent DIF> element in the DIF metadata standard supports linking between a dataset and the collection where the dataset belongs, thus creating a multi-level description. For example, the dataset 'Near-Surface Water-Quality Surveys of the Caloosahatchee River and Downstream Estuaries, Florida, USA' is a member of the collection 'USGS_SOFIA_Ding_Darling_baseline'. This information is recorded in the Parent DIF element of the metadata record of the dataset.

Entity linking and authority control

The national archives link the descriptions of records and the descriptions of other kinds of entities, which provide context to records. For example, at NARA users can navigate back and forth between archival description and authority records for government agencies, which provide more detailed information for record producers. In the catalogue of NAA, users are able to navigate back and forth between three types of entities: records, agency

and functions. Entity linking is also found in dataset repositories. GCMD maintains three separate catalogues, each for datasets, service/tools and ancillaries (projects, data centres, platforms and instruments). The dataset catalogue is linked to the catalogue for ancillaries. Ancillary descriptions provide background information related to datasets. In a sense, ancillary descriptions are like the authority records for archival creators in the catalogues of national archives. A user is able to navigate to the ancillaries from the dataset catalogue and vice versa.

The four national archives use authority control in describing dataset creators, subjects, geographic locations and so on. For example, NARA uses an Organization Authority File, a Person Authority File, a Topical Subject Thesaurus, a Geographic Authority File, a Specific Records Type Thesaurus and a Program Area Thesaurus in its catalogue. Similarly, dataset repositories also commonly use controlled vocabularies. ICPSR uses a Subject Thesaurus and a Geographic Name Thesaurus in indexing subject terms and geographic coverage of datasets. The UK data archive uses the Humanities and Social Science Electronic Thesaurus for subject indexing. At GCMD, the name of the ancillaries and keywords used in subject indexing are under authority control. The LTER catalogue uses the Andrews Experimental Forest thesaurus for subject indexing, and a classification system to indicate the taxonomic coverage of datasets.[17]

Browsing structures

In addition to metadata-based catalogues, national archives also create browsing structures to facilitate the discovery of records. For example, at NARA, users can browse the record groups through topic-based clusters (http://www.archives.gov/research/alic/tools/record-group-clusters.html). At LAC, users can browse records based on type, topic and alphabetic order. Similarly, data repositories also provide browsing structures based on provenances, classification systems and other criteria. Government data portals allow users to browse datasets based on provenances, such as organisations that publish datasets. The Harvard Dataverse Network provides a hierarchical structure for browsing datasets from top to bottom levels, including network, sub-network, dataverse, collection study and file. The dataverse level is like the organisation level in CKAN. It represents a single organisation or scholar. At ORNL DAAC, a user is able to browse the complete list of datasets through a three-level hierarchical structure. The top-level categories are the types of research projects, such as field campaign and land validation. The second-level categories are specific research projects, for example the Boreal Ecosystem-Atmosphere Study is a research project under field campaign. The third level includes the specific datasets produced during each research project, for example the Boreal Ecosystem-Atmosphere Study produced 274 datasets, each of which consists of a number of data files and documentation files. ICPSR allows users to browse datasets through a classification system. The classification system includes 19 categories and numerous sub-categories. The LISS panel data archive allows users to browse datasets through a list of topics, which is essentially a simple classification system that organises datasets into a hierarchical structure of categories and sub-categories. A similar approach is also used by the UK government data portal data.gov.uk.

Differences between national archives and dataset repositories

In addition to the common information organisation methods and tools illustrated above, dataset repositories use more flexible structures and specialised metadata schemas in organising and describing datasets. They also support easier data access and reuse. These unique features demonstrate their stronger expertise in datasets management.

Dataset repositories support poly-hierarchical structures and dynamic collections

Although both national archives and dataset repositories support multi-level structure, dataset repositories are sometimes more flexible and support poly-hierarchical structures. For example, the DIF metadata used by GCMD allows a child metadata record to point to more than one parent metadata record.[18] The Dataverse Network software supports dynamic and virtual collections. A dynamic collection is created through a query that gathers studies into a collection based on matching criteria. A study might match the query selection criteria at one time, but not match the criteria for that collection at another time owing to changes in the matching criteria. Virtual collections are created through linking. Users can link a collection from one dataverse to another dataverse.

Dataset repositories use metadata schemas specialised for datasets

The four national archives describe datasets following general archival description standards that apply to all types of archival materials. Their catalogues do not contain metadata elements specifically defined for datasets. For example, the Life Cycle Data requirements Guide,[19] which NARA uses for archival description, does not define any elements specific for datasets. Government data portals use metadata schemas for datasets in general, not specific to any type of dataset. The default metadata schema of CKAN, which is the common software platform for government data portals, includes one element defined specifically for datasets: Data Preview, which applies to any kind of data.[20] The metadata element set used by the Canadian government open data portal is based on Dublin Core.[21] The Australian government data portal uses the Common Core Metadata Schema.[22] The US government open data portal uses an updated version of the Common Core Metadata Schema, the Project Open Data Metadata Schema v1.1 (https://project-open-data.cio.gov/v1.1/schema/). Both the Common Core Metadata Schema and this updated version are based on DCAT.[23]

DCAT is an RDF vocabulary created by W3C for describing datasets. It is a general-purpose vocabulary and does not make any assumption about the format and type of the datasets. In other words, it describes datasets, but does not describe the special features of particular kinds of datasets, such as linked data, tabular data, network data, administrative datasets, research datasets, social science datasets or science datasets in particular domains. Other vocabularies may be used together with DCAT to provide more detailed description of particular kinds of data. The DCAT vocabulary defines a minimal set of classes and properties of its own and reuses terms from other vocabularies, such as Dublin Core, Friend of a Friend (FOAF) and Vcard. Its classes include: Catalog, Catalog Record, Dataset, Distribution (a specific form of a dataset for distribution, such as a csv data file), Concept Scheme (the knowledge organisation system used to represent themes/categories of datasets in the catalogue), Concept (a category or a theme used to describe datasets

in the catalogue) and Organization/person. The Catalog and Catalog Record classes and their properties are meta-metadata. They do not describe datasets directly but describe the catalogue of datasets.

CKAN, the software that powers the catalogues of the four government open data portals, also supports harvesting of external metadata, which are often created based on specialised metadata standards for particular types of datasets. In the UK, geospatial datasets are described using the GEMINI2 metadata standard and then both the datasets and their metadata are harvested into data.gov.uk.[24] In addition, data.gov harvests many metadata records encoded in ISO 19139 and the Federal Geographic Data Committee (FGDC), which are standards for geospatial datasets.

The four social science data archives all use DDI, a documentation standard created specifically for social science data. As a specialised metadata standard, DDI supports the description of unique features of social science research data, such as principle investigators, study, funding, universe, sample, data collection methodology, data processing procedures, and versioning of data and metadata. Over the years, DDI has been updated several times and has become complicated. The current version, DDI Life Cycle 3.2, defines 42 XML schemas and 1181 elements.[25] Although complicated, not all of the elements have to be implemented by a particular institution. An institution can choose a subset of name spaces and elements of DDI for local use. Each of the four social science data archives uses a subset of different versions of DDI. For example, the LISS panel data archive uses DDI 3, whereas Harvard Dataverse Network uses DDI 2.0. Harvard Dataverse Network uses only two of the seven main DDI sections, the fileDscr and dataDscr sections. Inside these two sections, only metadata elements that have direct equivalents in the Dataverse Network are supported.[26]

Unlike social science data archives that have a common metadata standard across domains, different scientific data repositories use different metadata standards created specifically for datasets in one or more scientific field(s). LTER uses the Ecological Metadata Language (EML) format. EML is a metadata standard developed for documenting research datasets in the earth, environmental and ecological sciences.[27] Similar to DDI, this metadata standard is defined in a modular and extensible manner, which means a particular user can select a number of modules to use and can create new modules when necessary. Each module contains a number of metadata elements describing a particular entity related to the dataset being described. These modules cover every aspect of datasets, including context, the dataset itself, content and the technical features of datasets exported from relational databases. The project module describes the research context in which the dataset was created. The dataset module provides overview information about datasets. The software module documents software needed in order to view or to process a dataset. The party module can be used to describe the creators of the datasets. The methods module describes the methods used in creating the dataset, including description of field, laboratory and processing steps, sampling methods and units, and quality control procedures. Three modules, entity, attribute and constraint, describe the internal components of a dataset. The entity module describes entities in the dataset, which are usually data tables. The attribute module describes variables in data tables. The constraint module defines the integrity constraints between entities (for example, data tables) as they would be maintained in a relational management system, such as primary key and foreign key constraints. The view module describes a view from a database management system.[28] The stored procedure module describes coded complex queries and transactions that can be invoked to produce data output from databases.

The catalogue of ORNL DAAC describes datasets using the Mercury 21.dtd metadata schema, which contains a subset of the FGDC metadata standard Content Standard for Digital Geospatial Metadata.[29] The metadata application profile of Dryad consists of 18 elements, which are drawn from the following metadata schemas: the Bibliographic Ontology (http://bibliontology.com/), dcterms (http://dublincore.org/documents/dcmi-terms/), Dryad Repository (http://datadryad.org/metadata/) and Darwin Core terms (http://rs.tdwg.org/dwc/index.htm).[30] The metadata application profile describes three types of entities: publications associated with the data packages preserved in Dryad, data packages and data files in the data packages.[31]

Specialised metadata schemas are able to describe the unique features of datasets and support searches based on these special features. For example, where geospatial information is provided, CKAN allows users to filter search results by geographical locations and to specify a bounding box to limit the area that users are interested in.[32]

Data repositories support easier access and reuse of datasets

Some national archives, such as NAA and LAC, do not provide data and documentation files online. Other national archives, such as NARA and TNA, provide some of their data and documentation files online. Dataset repositories usually provide data and documentation files online, unless there are access restrictions owing to privacy, embargo or other concerns. The data files at national archives are provided in various kinds of formats. Here is a list of data files from NARA: RG122.SAT.COACHED, RG137.FEDPROC.Y80, HEIDY.Y6872.DAT and HMS.CENS1998.zip. Notes are provided explaining the MIME types of the data files, such as text/plain or application/zip, which help decide the kinds of software that can open the data files. Documentation files provided by national archives are usually in PDF or other text formats. While these data and documentation file formats are usable in many circumstances, dataset repositories provide data and documentation files in formats that are easier to reuse and ready for statistical analyses. At ICPSR, each data file can be downloaded in SAS, SPSS, SATA and ASCII format. Each of the first three formats is ready to use in a particular kind of statistical software. If a user downloaded the data in ASCII format, they can also download setup files for a particular kind of statistical software so that the ASCII data file can be imported into the software for analysis. The Harvard Dataverse Network allows users to download data files in text format, R data, S plus and Stata format. It also allows users to download a subset of the variables or observations of tabular data files or a subset of vertices or edges of network data. In the LTER catalogue, the Code Generation element generates codes for analysing a data package in Matlab, R, SAS and SPSS, and provides instructions for how to run the codes in these kinds of statistical software.

Dataset repositories also provide online data visualisation and statistical analysis tools. The UK data archive provides the data analysis tool Nesstar online and allows users to view variable frequencies and conduct simple online tabulations and graphs.[33] ICPSR allows users to do similar things through its online data analysis tool Survey Documentation and Analysis. Where geospatial information is provided, CKAN can plot the data on an interactive map.[34] With geocoded data, GCMD displays the geographic coverage of datasets on a map. The service/tools catalogue of GCMD contains many data visualisation and analysis tools, such as the 'CanVis – Visualization Program For Seeing Potential Impacts from Coastal Development or Sea Level Rise'.

Metadata-based catalogues and browsing structures support the discovery of datasets. In addition, national archives and dataset repositories support the discovery of information within data files. But they support the discovery of different kinds of information from data files. For some of their datasets, national archives allow users to search for individual records (rows within data tables) within data files. For example, in the data file 'Russians to America Passenger Data File, 1834–1897' of the NARA Access to Archival Databases (http://www.archives.gov/aad), a user is able to search for passengers based on last name, first name, age, country of origin, destination city/county and so on. This kind of information discovery is useful for historians who are interested in historical facts. All the four social science data archives support the discovery of variables, that is columns in data tables. For survey data gathered through questionnaires, the LISS panel data archive and the UK data archive also support the discovery of individual survey questions. Searching for variables and survey questions is very useful for social science researchers who want to analyse and reuse data. National archives and dataset repositories support the discovery of different kinds of information probably because their designated communities vary. People use national archives to learn the history of a country, hence they may be interested in individual records. In contrast, people use dataset repositories to find the right data to analyse and reuse, therefore they may be more interested in statistical patterns rather than individual data points.

Dataset repositories tend to preserve multiple versions of datasets. For example, the LTER catalogue tracks different versions of the same data package. Each version is catalogued separately and the relationships between different versions are recorded and displayed. In the LTER catalogue, metadata records for derived data packages contain provenance metadata, which shows the title, creator, distribution and contact information of the source data package. Tracking version history helps users decide the authenticity of datasets and select the right version to use.

Some dataset repositories, such as data.gov and ICPSR, maintain download or view statistics for each dataset. Download or view statistics are an indicator of the popularity and impact of datasets. They are helpful for secondary data users in selecting datasets.

Conclusions

It is evident from the findings that dataset repositories provide more specialised description, and support deeper discovery and easier access and use for datasets, than national archives. In fact, national archives acknowledge their limitations in facilitating data reuse. As stated by an archivist from one of the four national archives:

> In the case of datasets, it is to be noted that we do not arrange or organize the data in order to make them easily understandable or easier to manipulate by the researchers. It is a task we cannot undertake because of our limited resources. It is up to the researchers to interpret and organize the data as they wish or can by using the metadata information provided. It is clear that our procedure for providing access to electronic documents needs to be improved. [We are] aware of this situation, but the lack of resources and other more urgent priorities make it difficult for the moment to find a permanent solution to this problem.[35]

There are some possible approaches that archival institutions can take to improve this situation. They may adopt some knowledge organisation and information discovery tools utilised by dataset repositories, such as specialised metadata and documentation standards

for datasets, data and documentation file formats that are ready for statistical analyses and machine processing, and online data visualisation and analysis tools. External funding can be applied to make this possible. For example, with the funding from National Digital Information Infrastructure and Preservation Program, the Electronic Records Custodial Division of NARA and four social science data archives formed the Data-PASS alliance, which created a union catalogue for the holdings of all partners.[36] The union catalogue is now accessible through the Harvard Dataverse Network. As mentioned earlier, the Harvard Dataverse Network offers advanced data discovery and analysis functionalities and uses DDI as the underlying metadata schema. Although currently the advanced functionalities of Dataverse are not fully utilised for NARA datasets,[37] collaboration with social science data archives is a step toward using specialised tools and metadata schemas for datasets.

Archival institutions might also consider releasing the management of datasets to dataset repositories. Archival institutions can still maintain the legal custody of datasets where necessary. In fact, this is what has been done by some archival institutions. For example, before 2010, TNA relied on NDAD in managing government datasets. Since 2010 when NDAD was discontinued, TNA has relied on the UK Government Web Archive to harvest data from government websites. TNA does not describe and organise the harvested datasets. Instead, it catalogues the archived websites and makes them accessible through the Discovery catalogue. The organisation and description of government-produced datasets depend on government agencies. In the United States, it has been NARA's policy to leave scientific data to data centres maintained by government agencies that possess the expertise for managing those data.[38] However, NARA still manages and preserves many government administrative datasets.

Endnotes

1. Society of American Archivists, 'Describing Archives: A Content Standard (DACS), Second Edition, 2013', available at <*http://files.archivists.org/pubs/DACS2E-2013.pdf*>, accessed 3 December 2014.
2. ibid., p. 142.
3. US Geological Survey, 'Standards', available at <*http://www.usgs.gov/core_science_systems/csas/metadata/standards.html*>, accessed 17 June 2015.
4. E Shepherd and C Smith, 'The Application of ISAD(G) to the Description of Archival Datasets', *Journal of the Society of Archivists*, vol. 21, no. 1, 2000, pp. 55–86.
5. The National Archives, 'Archiving Datasets', available at <*http://www.nationalarchives.gov.uk/webarchive/archiving-datasets.htm*>, accessed 3 December 2014.
6. Datasets originally preserved in NDAD are still accessible through the Discovery catalogue of TNA. For example, the 'Internal Drainage Board Database' can be found from both NDAD and the Discovery catalogue of TNA. Metadata for datasets in the TNA Discovery catalogue are mapped from NDAD. Some metadata from NDAD, such as table catalogues and documentation catalogues, is included as documentation files in the Discovery catalogue of TNA. Other metadata is mapped to some fields in the TNA catalogue. Sometimes the mapping is not appropriate. For example, these metadata elements in the NDAD metadata schema 'original system attributes', 'Hardware', 'Operating System', 'Application Software', as well as 'Logical structure and schema', were mapped to the 'Arrangement' element in the Discovery catalogue.
7. Cited from Shepherd and Smith. Originally from National Digital Archive of Datasets publicity leaflet, which is no longer accessible.
8. Shepherd and Smith, p. 57.

9. LTER Network, 'Data Access Policy, Data Access Requirements, and General Data Use Agreement', available at <*http://www.lternet.edu/policies/data-access*>, accessed 3 December 2014.
10. Simon Demissie, 'The UK Government Web Archive: A Resource for Contemporary Historians', available at <*http://blog.nationalarchives.gov.uk/blog/uk-government-web-archive-resource-contemporary-historians-2/*>, accessed 3 December 2014.
11. Dataverse project, 'User Guide', available at <*http://guides.dataverse.org/en/4.0.1/user/*>, accessed 3 December 2014.
12. 'Harvard Dataverse Network', available at <*https://thedata.harvard.edu/dvn/*>, accessed 3 December 2014.
13. 'About ORNL DAAC', available at <*https://daac.ornl.gov/about_us.shtml*>, accessed 24 June 2015.
14. 'Data Policy Statements', available at <*http://www.data.gov/data-policy*>, accessed 3 December 2014.
15. Dryad Development Team, 'Dryad Frequently Asked Questions', available at <*http://datadryad.org/pages/faq#deposit*>, accessed 5 December 2014.
16. DDI Alliance, 'DDI 3.2 XML Schema Documentation', available at <*http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/*>, accessed 4 December 2014.
17. LTER Network, 'LTER Controlled Vocabulary', available at <*http://vocab.lternet.edu/vocab/vocab/index.php*>, accessed 2 November 2014.
18. Global Change Master Directory, 'Parent DIF', available at <*http://gcmd.nasa.gov/add/difguide/parent_dif.html*>, accessed 5 December 2014.
19. NARA, 'Lifecycle Data Requirements Guide', 2002, available at <*http://www.archives.gov/research/search/lcdrg/*>, accessed 4 December 2014.
20. CKAN, 'Publish and Manage Data', available at <*http://ckan.org/features/#publish*>, accessed 4 December 2014.
21. Government of Canada, 'Open Data Metadata Element Set', available at <*http://open.canada.ca/en/metadata-element-set*>, accessed 4 December 2014.
22. Australian Government, 'Open Data Toolkit', available at <*https://toolkit.data.gov.au/index.php?title=Definitions#Common_Core_Metadata*>, accessed 4 December 2014.
23. W3C, 'Data Catalog Vocabulary (DCAT)', 2014, available at <*http://www.w3.org/TR/vocab-dcat/*>, accessed 4 December 2014.
24. 'Getting Started: Initial Guidance to Data Providers and Publishers', 2011, available at <*http://data.gov.uk/sites/default/files/UKL-Getting-Started-Guide-0-v2-0_10.pdf*>, accessed 4 December 2014.
25. DDI Alliance.
26. Harvard Dataverse Network, 'User Guide', available at <*http://thedata.harvard.edu/guides/dataverse-user-main.html#ddixml-datafile-ingest*>, accessed 4 December 2014.
27. Knowledge Network for Biocomplexity, 'Ecological Metadata Language (EML) Specification', available at <*https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html#eml-view*>, accessed 5 December 2014.
28. A view provides filtered slices of a database for targeted analysis. In other words, it selects a number of data elements from a number of data tables in a database and present to users for a particular purpose.
29. DataONE, 'Mercury Metadata Editor', available at <*https://www.dataone.org/software-tools/mercury-metadata-editor*>, accessed 5 December 2014.
30. Dryad Development Team, 'Dryad Metadata Application Profile', Version 3.0. 2010, available at <*http://wiki.datadryad.org/wg/dryad/images/8/8b/Dryad3.0.pdf*>, accessed 5 December 2014.
31. Dryad Development Team, 'Dryad Frequently Asked Questions'.
32. CKAN.
33. UK Data Archive, 'Online Data Browsing', available at <*http://www.data-archive.ac.uk/find/online-data-browsing*>, accessed 17 June 2015.
34. CKAN.

35. This quote is from the answer to a reference question sent to one of the national archives.
36. M Altman, M Adams, J Crabtree, D Donakowski, M Maynard, A Pienta and C Young. 'Digital Preservation Through Archival Collaboration: The Data Preservation Alliance for the Social Sciences', *The American Archivist*, vol. 72, no. 1, Spring/Summer 2009, pp. 170–84.
37. This finding is based an examination of NARA data at the Harvard Dataverse Network on 1 March 2016.
38. NARA, 'Strategic Directions: Appraisal Policy', 2007, available at <*http://www.archives.gov/records-mgmt/initiatives/appraisal.html*>, accessed 25 June 2015.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Jinfang Niu is an Assistant Professor at the School of Information, University of South Florida. She received her PhD from the University of Michigan, Ann Arbor. Prior to that, she worked as a librarian at the Tsinghua University Library for three years. Her current research focuses on electronic records and digital curation.