Data appraisal and Ingest

Ingest is the process of transferring data into an archive or repository for long-term preservation and, often, dissemination. Specifically, this concept considers the role of the archive between the end of the initial research and the release of data for re-use. The ingest process sees quality checks performed on the data and documentation. If not already done, metadata is added that is relevant to the discovery, preservation, and re-use of the data (see Documentation and Metadata section). In addition, where necessary, data is converted to the format(s) used for archiving and dissemination. Communication between data creator and archive is vital in data ingest, particularly agreement on data transfer in accepted file formats and clarifying the legal status of data.

Data appraisal, or what to archive?

Archives do not operate in isolation. They exist to serve a community of users and producers of data. Neither can they archive everything. Therefore, an archive needs to define the range of its collection and what type of data and formats it intends to curate. For example, is the archive discipline specific? GESIS, for instance, is a social and political science data archive and is only interested in data that falls within those disciplines[1]. Does the archive concentrate on a particular spatial unit? National, regional, or local archives are interested in data created in or specific to that locality. Is the archive interested in specific formats? The British Library's Sound Archive[2], which curates sound and visual data, is an example. Will the archive accept machine-readable formats only? Alternatively, is it a physical archive with artifacts requiring researchers travel to the archive to use?

Here it is important to highlight the distinction between re-use and preservation. The terms are not synonyms. Making data available for re-use does not necessarily imply a long-term commitment to preservation; likewise, a long-term commitment to preservation does not imply data is available for re-use. Before accepting data, an archive therefore needs to clarify its objectives and policy concerning these issues, and address them in its communication with data creators and depositors.

Resolving the question of what not comes partially from legal issues. Archives cannot accept material without permission to archive from the intellectual property right holder. Ethical issues are additional selection criteria: archives cannot accept material where consent forms and verbal consent agreements have explicitly prohibited archiving and/or data re-use. In addition, if there is a case to be made for an embargo on the data, is that embargo reasonable? If, for example, the data is likely to become quickly redundant then the case for archiving that data is weak

[1] http://www.gesis.org/en/services/archiving-and-registering/data-archiving/
[2] http://sounds.bl.uk/

CESSDA AS
CESSDA House

Parkveien 20
5007 Bergen
NORWAY

phone: +47 55 58 21 18
e-mail: cessda@cessda.net
www.cessda.net

Archives have to make judgments on the value of preservation. Simply, are the data worth preserving, for re-use or otherwise? Both avenues involve significant commitments in time and resources. Consequently, archives should employ a set of criteria to support the selection process and decision.

Some criteria to consider are

- The relevance, uniqueness, and value of data for research.
- The quality of data and documentation: is the data of a reasonable standard, is its documentation sufficient to ensure it is understandable, and is viable for re-use?
- Technical aspects: Can the archive handle the format in which the data is offered?
- What are the expected costs for preserving and making the data available for re-use?
- Does the data fit into the scope of your collection? Would another organization or institution be better suited to the data?

In order to define a set of criteria meaningful and relevant to your archive, it is helpful to first assess which data are created by the researchers who are your main data producers, and what their attitudes to data management and data sharing are. A tool supporting this assessment is the Data Asset Framework[3] (DAF). While aimed primarily at Higher Education Institutions (HEI), DAF can also help archives from non-HEI contexts to explore how their main target group produces, manages, stores, and shares data.

Once adopted, a data acquisitions policy or mandate will shape the ingest process. If you are operating on a fixed agreement or mandate for data ingest, then talking to data producers can help manage the process better by seeing if they can deposit data in agreed file formats. If they cannot agree, then consider if it is suitable or feasible to set up ingest training? However, if the basis of your acquisitions process is voluntary depositing of data then the ability to shape good quality data into ingest-ready formats is restricted.

Questions to consider in relation to file formats include:

- Can your ingest process handle material in any file format or is it restricted in what formats it can accept?
- How to handle file formats? For example, will they be retained in their native form or migrated to a standard preservation or dissemination format.
- While migrating to a regularized format may be advantageous to the archive, what is its impact on the functionality of data?
- Would such a strategy bring authenticity issues (e.g. are users happy to work with manifestations of, rather than original, data formats?)
- How accessible are regularized formats?

Examples:

- DANS: Preferred Formats http://www.dans.knaw.nl/sites/default/files/file/EASY/DANS%20preferred%20formats%20UK%20DEF.pdf

[3] http://www.data-audit.eu/

- GESIS Data Archive: Recommended formats http://www.gesis.org/en/services/archiving-and-registering/data-archiving/preparing-data-for-submission/

Recommended introductory resources

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of digital social science data. *Data Science Journal*, *3*(2004), 209-221. doi:http://dx.doi.org/10.2481/dsj.3.20

Tjalsma, H., & Rombouts, J. (2010). *Selection of Research Data. Guidelines for Appraising and Selecting Research Data*. Retrieved from https://www.surf.nl/en/knowledge-and-innovation/knowledge-base/2010/research-report-selection-of-research-data.html

Whyte, A., & Wilson, A. (2010). *How to Appraise & Select Research Data for Curation. A Digital Curation Centre and Australian National Data Service "working level" guide.* Retrieved from http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Example policy

UK Data Service. (2014). Collections Development Selection and Appraisal Criteria. Retrieved from http://ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf

Deposit agreements

To accept data into an archive, archives require depositors sign a submission or deposit agreement (also called a license agreement) to ensure archives have the right to manage long-term preservation and re-use. Preparing and managing long-term digital preservation requires archives migrate and manipulate data, so they must make and retain multiple copies of data and documentation. However necessary, this involves altering the data in a way that can infringe intellectual property rights. Consequently, a legal contract must exist between depositor and archive clearly stating what actions archives can perform to data. In addition to regulating conditions under which data can be re-used, agreements must grant archives rights to copy data for bitstream preservation, to modify for preservation (for instance, migrating to a different format), back up, and to repackage and enhance and to enable data discovery and facilitate re-use.

Examples for deposit agreements from CESSDA archives (English language versions):

Czech Republic - CSDA	http://archiv.soc.cas.cz/sites/default/files/agreement_on_data_deposition_en.doc
Denmark - DDA	http://samfund.dda.dk/dda/Depositorform.doc
Finland - FSD	http://www.fsd.uta.fi/en/forms/deposit.pdf
France - Réseau Quetelet	http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=3&lang=en
Germany - GESIS	http://www.gesis.org...Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf
Netherlands - DANS	http://www.dans.knaw.nl/en/content/dans-licence-agreement-deposited-data
Norway - NSD	http://www.nsd.uib.no/nsddata/arkivering/en/006_archiving_agreement.pdf
Sweden - SND	http://snd.gu.se/en/deposit-data/deposit-agreement
Switzerland - FORS	http://forscenter.ch/wp-content/uploads/2013/11/deposit_contract_e.pdf
United Kingdom - UK Data Archive	http://ukdataservice.ac.uk/media/28102/licenceform.pdf

Quality assurance

After data has been submitted to the archive, a number of quality checks and other measures to assure quality are carried out, concerning both technical and content-related aspects. On the content side, quality checks should focus on both data and documentation, for example, asking if documentation is sufficient to guarantee future users will be able to understand the data? Is data consistent? Any confidentiality issues that need resolving before the data can be archived?

Checks also make sure ingested data is suitable for preservation and dissemination by establishing the integrity, "completeness" and "intactness" of submitted files (see nestor, 2009). On a basic level this involves quarantine and checking of files for malware, cyclic redundancy checks, as well as format identification (determining the file format of the digital object) and validation (determining whether the file indeed meets format specifications; see http://jhove.sourceforge.net/).

Support for these processes comes from various tools as well as file format registries. Some examples:

- PRONOM, developed by the UK National Archive: http://apps.nationalarchives.gov.uk/PRONOM/.
- The Unified Digital Formats Registry (UDFR) developed by the University of California Curation Center as part of the National Digital Information Infrastructure Preservation Program in the U.S.: http://www.udfr.org/.
- An extensive, up-to-date lists of digital preservation tools (including tools for ingest) is available from COPTR, the Community Owned digital Preservation Tool Registry: http://coptr.digipres.org/.

References

nestor Working Group Trusted Repositories - Certification. (2009). nestor criteria. Catalogue of Criteria for Trusted Digital Repositories. Version 2. Retrieved from http://files.d-nb.de/nestor/materialien/nestor_mat_08_eng.pdf

Further Reading

CCSDS. (2004). Producer-Archive Interface Methodology Abstract Standard. Retrieved from http://public.ccsds.org/publications/archive/651x0m1.pdf

nestor working group long-term preservation standards. (2009). Into the Archive: A Guide for the Information Transfer to a Digital Repository. Draft for public comment. Retrieved from http://files.d-nb.de/nestor/materialien/nestor_mat_10_en.pdf

Acquisition policy checklist

	Yes	Partly	No	Notes
Does your organization have a defined mission statement?	●	●	●	The mission statement determines the scope of the collection (usually in very general terms)
Are you aware of all legal requirements determining which material has to be preserved or cannot be accepted by your archive or repository?	●	●	●	Funding mandates, data protection laws, Freedom of Information requirements.
Has your organization defined its designated community, now and in the future?	●	●	●	It's important to know who is offering objects, who wants to use them, how, and in what ways? This determines what you accept and reject.
Did your organization assess its core areas of expertise?	●	●	●	What can you cope with? If there is no expertise in handling audiovisual formats, the Archive cannot accept these materials
Are you familiar with the collection profiles of other archives in your subject field?	●	●	●	If a certain type of content isn't preserved anywhere else, it might be a reason to consider it for your archive or repository. Alternatively, if certain materials are already covered by other archives, there may be no need for you to archive them
Have you defined concrete criteria to guide the act of appraisal?	●	●	●	Do you have a written policy and guidelines, or a checklist to match against submissions?
Do these criteria consider • content • format • quality • uniqueness/historical value • relevance • economic value?	●	●	●	