Irish Social Science Data Archive

Data Acquisition Protocol

The Irish Social Science Data Archive (ISSDA) is Ireland's centre for quantitative data acquisition, preservation, and dissemination.

Based at UCD Library, its mission is to ensure wide access to quantitative datasets in the social sciences, and to advance the promotion of international comparative studies of the Irish economy and Irish society.

Purpose

The ISSDA Data Acquisition Protocol specifies procedures and workflows to be followed when a Data Study has been received from a Data Depositor. The data and supporting documentation supplied by a Data Depositor comprises a Submission Information Package (SIP). The ISSDA Data Acquisition Protocol outlines the process for validating the SIP and creation of an Archival Information Package (AIP) and Dissemination Information Package (DIP) managed within archive storage and, in most cases, the NESSTAR data access system.

Overall Framework

ISSDA has broadly adopted the Open Archival Information System (OAIS)[1] reference model as a framework on which our workflow is based. An OAIS is an archive that has accepted the responsibility to preserve information and make it available for a Designated Community. ISSDA's Designated Community consists of educators and researchers across multiple sectors and jurisdictions with broad relevance to the Social Sciences and Public Health.

As a reference model, the OAIS's primary purpose is to provide a common set of concepts and definitions that can facilitate the specification of archives and digital preservation systems. Activities or functions in an OAIS include ingest, preservation planning, administration, data management, archival storage and data access.

[1] Definition adapted from the Digital Preservation Handbook, 2nd Edition, http://handbook.dpconline.org/ Digital Preservation Coalition © 2015 licensed under the Open Government Licence v3.0

March 2017

OAIS was first approved as ISO Standard 14721 in 2002 and a 2nd edition was published in 2012. Although produced under the leadership of the Consultative Committee for Space Data Systems (CCSDS), it had major input from libraries and archives.

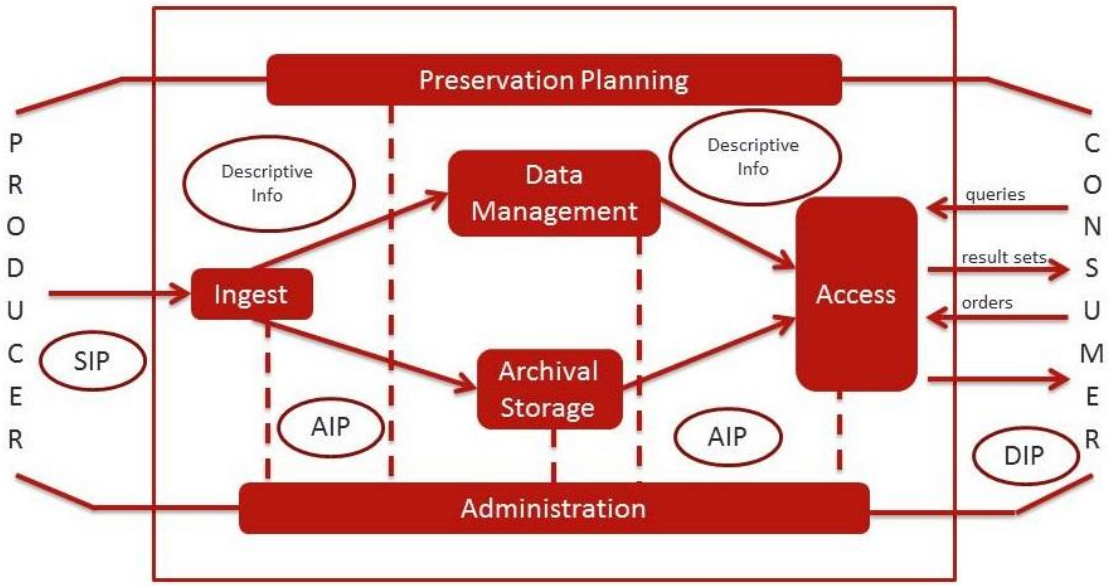

Figure 1: OAIS reference model

Pre-Ingest

Although 'Pre-Ingest' is not part of the OAIS model it is a very valuable part of ISSDA's workflow. ISSDA works with individual depositors to ensure data are deposited with sufficient metadata, documentation and in a preferred file format, so that they can be fully understood and re-used by others. All studies offered for submission to ISSDA are appraised in accordance with the criteria outlined the Collection Development Policy[2]. Information on preferred file formats is available in ISSDA's File Format Policy[3].

We advise Depositors on ensuring the data are cleaned and on provision of suitable documentation, including a data dictionary or codebook and questionnaires, as well as preservation friendly formats for both data and documentation. We can also point Depositors to useful information on the anonymisation process for quantitative data.

Such pre-ingest activity helps to ensure that issues that may affect the quality of data deposited, for example legal, ethical and rights management issues together with file formats for dissemination and curation, are considered and addressed early in the process.

[2] http://www.ucd.ie/t4cms/ISSDA_Collection_Development_Policy_V1.pdf
[3] http://www.ucd.ie/t4cms/ISSDA_Format_Policy_V1.pdf

Ingest

Ingest is the process of putting data into a digital archive, in this case ISSDA. This includes the receipt of data, documentation, Depositor Form[4] and Deposit Licence[5] from the Data Depositor. These 'original' files are the Submission Information Package (SIP) and are preserved in the original format in the appropriate directory of the archival storage. During the ingest process the SIP is verified and validated, data are checked, Archival Information Package (AIP) and Dissemination Information Packages (DIP) are generated and descriptive metadata for the Study are created.

- Data delivery: Discuss with Data Depositor the best way to transfer data. Possible options include encrypted files sent via email, with encryption password sent using a different medium, FTP transfer or HEAnet Filesender Guest Voucher[6].
- Receipt of data: Upon receipt of the SIP a Study Number (SN) is assigned to the Study. This is in the format XXXX-XX, e.g. 1234-01. Where the Study is a one off Study the Study Number ends in -00; where the Study will have multiple waves of data the Study Number ends in -01, -02 etc. with the number indicating the number of the wave.
- Data verification: For all files, both the data and all accompanying documentation, ISSDA requests that the depositor indicate filename, file format and a description of the file contents in the Depositor Form. This is considered to be the file manifest. This allows the Data Manager to verify that the SIP is complete; that all files are present, that they conform to the declared file format and can be opened in associated software applications.
- Data validation: If a checksum is delivered with the SIP this is recorded and validated to ensure the integrity of a file and verify it has not been altered or corrupted en route to ISSDA. If one is not delivered a MD5 checksum is generated once the data are in archival storage using a tool called Fixity[7] from AVPreserve to be used for future fixity checks.
- Data checking: All data and documentation, together with the deposit form, are sent to the Centre for Support and Training in Analysis and Research (CSTAR)[8] for data checking. CSTAR inspect the quality of the data, assuring that the data is in a format suitable for analysis and that the data is clean, with variable level codes, missing value codes, descriptive and consistent variable names, etc. They additionally check that there are no variables considered direct identifiers and evaluate related accompanying files to verify contextual information is given for the dataset provided. Once the data have been reviewed CSTAR send a report to the ISSDA identifying any concerns as outlined above, including specific actions which must be undertaken by the Data Depositor for correction. If changes need to be made as a result of CSTAR's recommendations the Depositor is given the opportunity resubmit the data to ISSDA. Subsequent versions of the data are not sent to CSTAR for checking.
- Metadata creation: ISSDA makes use of the NESSTAR Publisher[9] for creation of Data Documentation Initiative (DDI)[10] metadata; descriptive and variable-level metadata are

[4] http://www.ucd.ie/t4cms/ISSDA_Depositor_Form_V1.docx
[5] http://www.ucd.ie/t4cms/ISSDA_Deposit_Licence_V1.2.docx
[6] https://www.heanet.ie/services/hosting/filesender
[7] https://www.avpreserve.com/tools/fixity/
[8] http://www.cstar.ie/
[9] http://www.nesstar.com/software/publisher.html
[10] http://www.ddialliance.org/

generated from the Depositor Form and documentation deposited with ISSDA and made available via ISSDA's public data catalogue, supported by NESSTAR.

- Deposit processing: Processing of deposited data is minimal, as most issues are dealt with during the Pre-Ingest phase. However, it may be necessary, for example, to shorten variable lables to less than 60 characters before adding to NESSTAR.

Archival Storage

ISSDA is committed to maintaining deposited data for the long term. The Archival Storage function manages the digital objects (data and documentation) entrusted to ISSDA.

ISSDA's Archival Information Packages (AIP) contain a number of elements. These include:

- Depositor Form and Deposit Licence
- 'Original' files, as deposited to ISSDA. In the case of a new edition being deposited these are added to the AIP.
- A record of any deposit processing
- NESSTAR file
- DDI metadata in XML format
- CSTAR report
- Correspondence with the Depositor

Archival storage is monitored using fixity checks on both the primary storage and backup storage. ISSDA uses a tool called Fixity[11] from AVPreserve to perform this function. Checks are scheduled to run automatically, once a week, and results are emailed to the Data Manager.

Data Management

The Data Management function works in conjunction with the Archival Storage function.

The Dissemination Information Package (DIP) is created as part of the Ingest process. The DIP consists of the files that are sent to users once a data request has been accepted. All files (data and documentation) are given a new file name using the format SN_name_year, all lowercase e.g.

- 0021-01_healthy_ireland_2015.dta
- 0036-01_qnhs_module_on_disability_Q2_2002.sav
- 0053-02_tilda_wave2_2012-2013.dta

Acronyms may be used if Study name is very long and version numbers are included where applicable, e.g 0015-12_eu-silc_2014_v2.

A cover page is inserted into deposited documentation, which outlines basic metadata for the Study including the Study Number (SN), title, depositor, publisher (ISSDA), the Study URL, publication date, version, type (dataset) and suggested citation. This is based on the metadata elements needed for

[11] https://www.avpreserve.com/tools/fixity/

data citation as outlined in Table 1, A Data Citation Roadmap for Scholarly Data Repositories[12]. This is to help users to identify the metadata relevant to the Study they are working with as well as providing a suggested citation for the data.

All Studies are disseminated using the same folder structure, allowing for easier navigation and identification of relevant files by users.

- DIP
    - Documentation
    - Data: If multiple formats, different folder for each
        - SPSS
        - Stata
        - etc.

Access

Once the Submission Information Package (SIP) is received and the Archival Information Package (AIP) and Dissemination Packages (DIP) are created, a record is added to the ISSDA website. This page contains DDI metadata for the Study including abstract, main topics, coverage, universe and methodology. It also outlines the data that are available, including the format(s) available, as well as linking to the available documentation provided by the Depositor.

To access the data users are asked to complete and sign an End User Licence (EUL)[13]. Separate licences cover use of data for research purposes or for use in teaching[14]. Data are supplied electronically using HEAnet Filesender[15]. All files are encrypted before being sent and a decryption password is sent to the user under separate cover.

ISSDA is in the process of upgrading our implementation of NESSTAR for creation of DDI metadata; descriptive and variable-level metadata are made available via ISSDA's public data catalogue, supported by NESSTAR. In 2017 ISSDA will begin to allow online access to data via NESSTAR[16]. This will allow users to find, browse, visualise and analyse data online. Users will still be required to apply for access to the data using the existing process but they would be able to browse the variable descriptions, frequencies, question text and other metadata in advance.

Roles and Responsibilities

Decisions within the workflows are ultimately made by the Data Manager in consultation with other staff members and the designated contact within the Data Depositor where necessary.

[12] Martin Fenner, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Robin Berjon, Sebastian Karcher, Maryann Martone, TimothyClark.
bioRxiv 097196; doi: https://doi.org/10.1101/097196
[13] http://www.ucd.ie/t4cms/ISSDA_Application_Research_V2.3.docx
[14] http://www.ucd.ie/t4cms/ISSDA_Application_Teaching_V2.1.docx
[15] https://www.heanet.ie/services/hosting/filesender
[16] http://www.ucd.ie/t4cms/ISSDA%20&%20NESSTAR.pdf

Review

This Protocol will be reviewed regularly in light of any relevant developments. Change management of the workflows is overseen by the Data Manager.

Appendix I: Inventory of Software Tools

The following is a list of software tools that ISSDA use:

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) Toolkit: DRAMBORA[17] is a toolkit for a digital repository audit. The toolkit guides users through the audit process, from defining the purpose and scope of the audit to identifying and addressing risks to the repository.

Fixity: ISSDA uses a using a tool called Fixity[18] from AVPreserve for fixity checking. By checking file fixity on ingest, or creating a checksum if one isn't delivered ISSDA is able to assert the authenticity of the data and documentation and establish a baseline fixity so unwanted data changes can be detected.

HEAnet FileSender: ISSDA uses FileSender[19] from HEAnet for dissemination of data. FileSender is developed to the requirements of the higher education and research community and supported by HEAnet, Ireland's National Education and Research Network.

Kleopatra: Kleopatra[20] is used by ISSDA to for decryption of datasets received from depositors. It is a certificate manager and a universal crypto GUI. It supports managing X.509 and OpenPGP certificates in the GpgSM keybox and retrieving certificates from LDAP servers.

PKWARE: ISSDA encrypts datasets before sending them electronically to users. All of the datasets are zipped (compressed) and encrypted using the commercial software, PKWARE[21]. PKWARE allows decryption using a pass-phrase. The files are sent using FileSender, a secure file sender service, which sends a message to the end user's email address. A separate email is sent to the user containing the pass-phrase.

NESSTAR: NESSTAR[22] is a software system for publishing data on the web. ISSDA makes use of the NESSTAR Publisher for creation of Data Documentation Initiative (DDI)[23] metadata and data discovery. Since January 2017 a project to re-catalogue of all datasets in ISSDA has been ongoing. Once the re-catalgouing is completed users will be able to perform simple online tabulations, produce graphs online without the need for specialist software and download the data in their preferred format.

UCD Connect: ISSDA uses the UCD staff email system. UCD email is provided in partnership with Google.

[17] http://www.repositoryaudit.eu/
[18] https://www.avpreserve.com/tools/fixity/
[19] https://www.heanet.ie/services/hosting/filesender
[20] https://www.kde.org/applications/utilities/kleopatra/
[21] https://www.pkware.com/
[22] http://www.nesstar.com/software/publisher.html
[23] http://www.ddialliance.org/

WinSCP: WinSCP (Windows Secure Copy)[24] is a free and open source SFTP, FTP, WebDAV and SCP client for MS Windows. Its main function is secure file transfer. Beyond this, WinSCP offers basic file manager and file synchronization functionality. For secure transfers, it uses Secure Shell (SSH) and supports the SCP protocol in addition to SFTP.

Glossary[25]

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an Archive (OAIS term).

Checksum: A unique numerical signature derived from a file. Used to compare copies.

Designated Community: An identified group of potential consumers who should be able to understand a particular set of information from an archive. These consumers may consist of multiple communities, are designated by the archive, and may change over time (OAIS term).

Dissemination Information Package (DIP): An Information Package, derived from one or more Archival Information Packages (AIPs), and sent by Archives to the Consumer in response to a request to the Archive (OAIS term).

Fixity Check: A method for ensuring the integrity of a file and verifying it has not been altered or corrupted. During transfer, an archive may run a fixity check to ensure a transmitted file has not been altered en route. Within the archive, fixity checking is used to ensure that digital files have not been altered or corrupted. It is most often accomplished by computing checksums such as MD5, SHA1 or SHA256 for a file and comparing them to a stored value.
http://en.wikipedia.org/wiki/File_Fixity

Ingest: The process of turning a Submission Information Package (SIP) into an Archival Information Package (AIP), i.e. putting data into a digital archive (OAIS term).

Open Archival Information System (OAIS): An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities, as defined in section 4 of the OAIS standard that allows an OAIS Archive to be distinguished from other uses of the term 'Archive'. The term 'Open' in OAIS is used to imply that the OAIS standards are developed in open forums, and it does not imply that access to the Archive is unrestricted. The OAIS abbreviation is also used commonly to refer to the Open Archival Information System reference model standard which defined the term. The standard is a conceptual framework describing the environment, functional components, and information objects associated with a system responsible for the long-term preservation. As a reference model, its primary purpose is to provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems. It has a very basic set of

[24] https://winscp.net/eng/download.php
[25] Definitions are taken from the Digital Preservation Handbook, 2nd Edition, http://handbook.dpconline.org/ Digital Preservation Coalition © 2015 licensed under the Open Government Licence v3.0

conformance requirements that should be seen as minimalist. OAIS was first approved as ISO Standard 14721 in 2002 and a 2nd edition was published in 2012. Although produced under the leadership of the Consultative Committee for Space Data Systems (CCSDS), it had major input from libraries and archives.

Submission Information Package (SIP): An Information Package that is delivered by the Producer to the Archive for use in the construction or update of one or more Archival Information Packages (AIPs) and/or the associated Descriptive Information (OAIS term).