Long-lived Data: Tools to Preserve Data

Theoretical Overview

Blue Coat Photos at
https://www.flickr.com/photos/111692634@N04/16042227002

Presentation by Johann van Wyk at the NeDICC Workshop on Data Preservation held on 15 February 2017, CSIR, Pretoria, South Africa

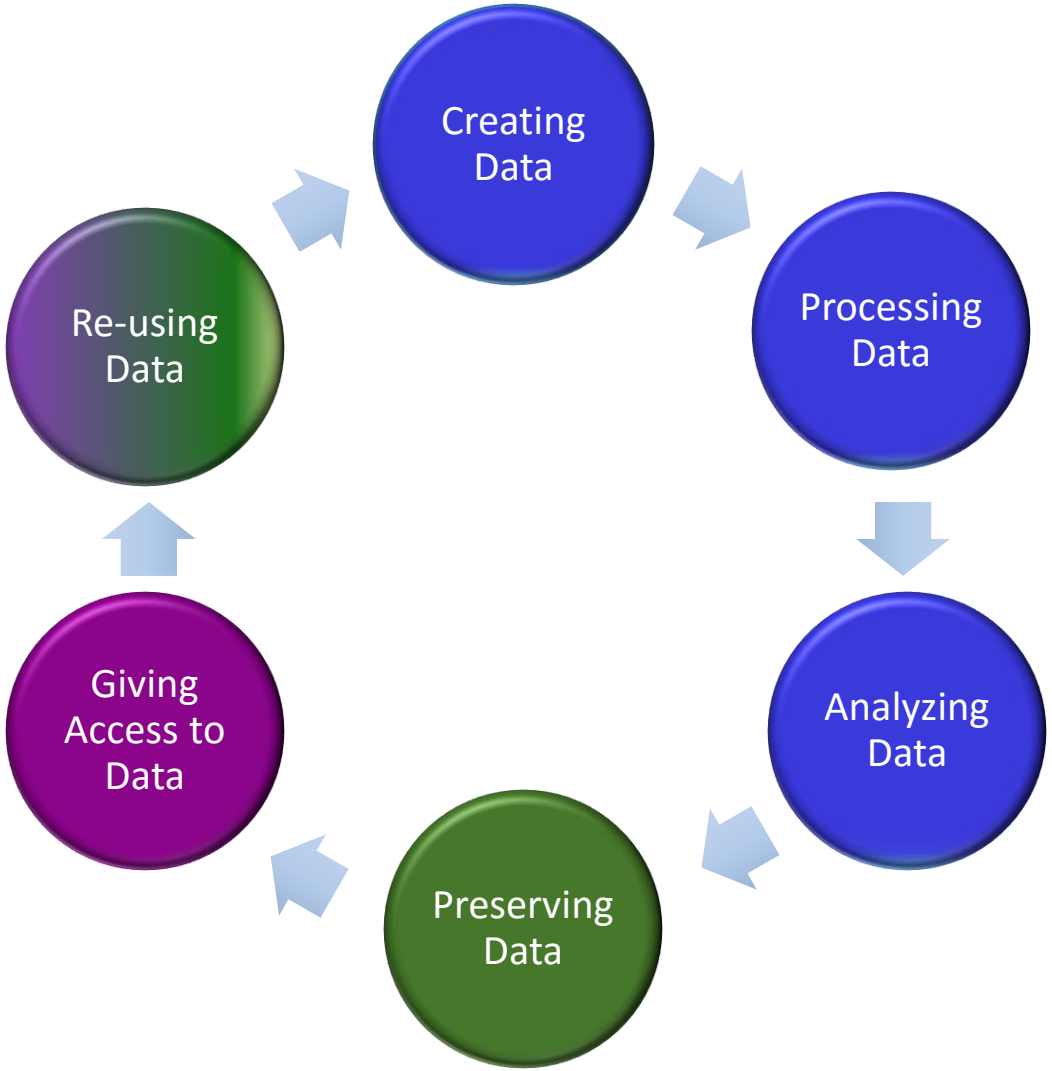
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

NeDICC
Network of Data & Information Curation Communities

Content

- Research Data Management Life-Cycle
- What is data preservation?
- Why should we preserve data?
- Taking steps to preserve data
- Metadata for preservation
- Levels of digital preservation
- What should be preserved?
- Institutional Readiness

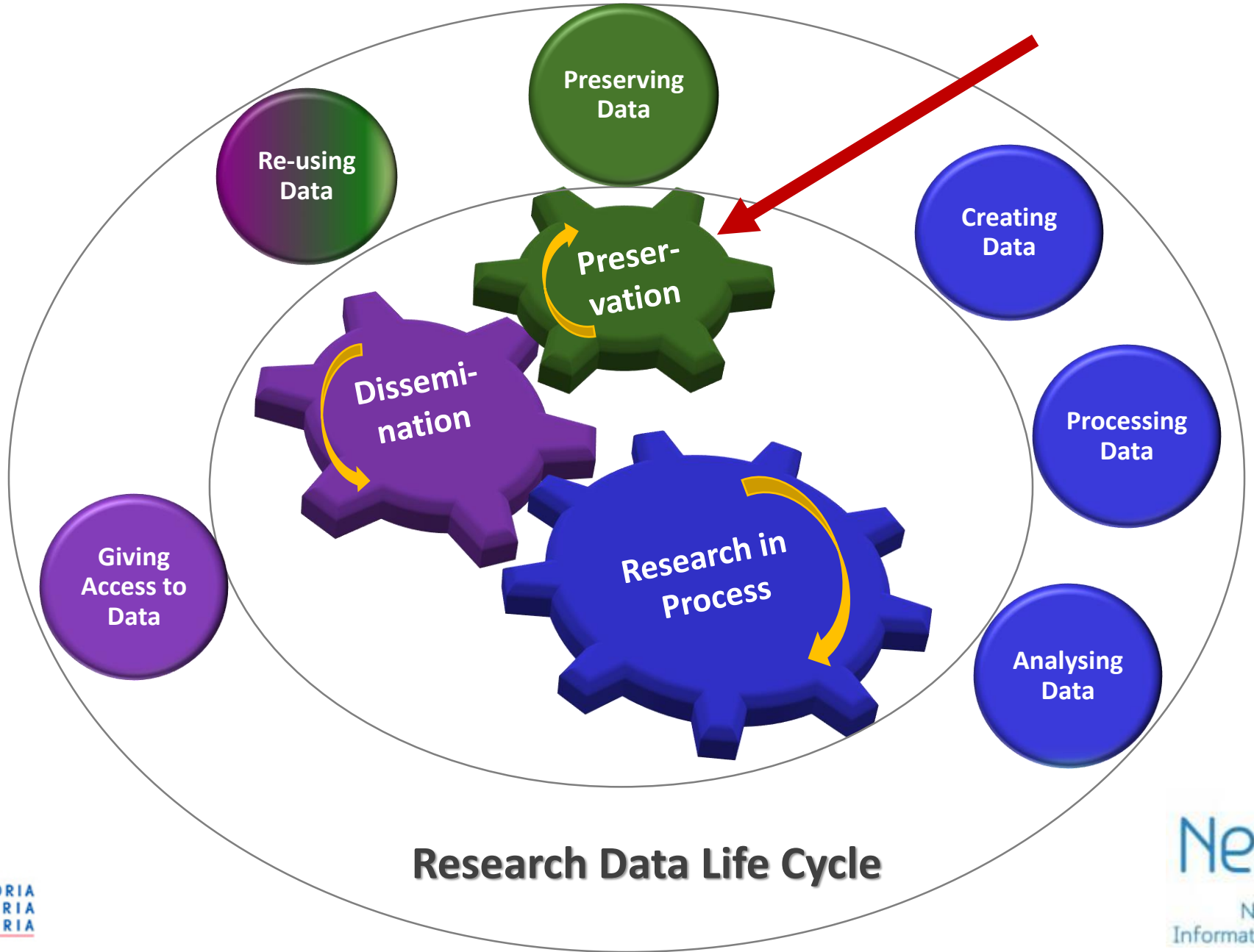
Research Data Life Cycle



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

NeDICC
Network of Data & Information Curation Communities

PROCESSES within the RESEARCH DATA LIFE CYCLE



What is Data Preservation?

- "Data preservation, or more specifically digital data preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary". This entails "all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change" (International Federation of Data Organizations for Social Science, 2012)
- Data preservation is "the process of providing enough representation information, context, metadata, fixity, etc. to the data and [thereby securing the data] so that anyone other than the original data creator can use and interpret the data" (Ruth Duerr, National Snow and Ice Data Center as cited by Choudhury, 2014)

Why should we preserve data?

Technical reasons:

- Lack of sustainable hardware, software, or support of computer environment may make the information inaccessible (hardware and software change over time) (Meghini, 2013: 51);
- File formats can become obsolete, for example:
  - The software or file formats are upgraded and the new version is not compatible with the old version;
  - "The software that supports the format is bought out by a competitor and withdrawn";
  - "The format falls into disuse", or no-one writes software that supports it;
  - The format is no-longer compatible with current software (ANDS Guide: file formats, 2016);

Ryan Somma at https://www.flickr.com/photos/ideonexus/6190987532

Why should we preserve data?

Technical reasons (Continued):

- Storage media can become corrupted due to decay (*The CESSDA User Guide on digital preser*vation, 2016).
- Compressed files might be more susceptible to "bit-rot", in other words there is risk that 'bits' of a data set might be changed during compression, causing changes throughout the entire document, rendering it useless (ANDS Guide: file formats, n.d.).
- Loss of the ability to identify the location of the data (Meghini, 2013: 51).

Why should we preserve data?

Content-related reasons:

- ❑ Users may be unable to understand or use the data, e.g.
  - researchers can develop their own methods to create/generate the data;
  - researchers can develop their own file naming convention, code books, terminologies;
  - data entered in rows and columns in an Excel spreadsheet, is only as useful as the headings or metadata that are supplied with it (Lebow and Carusso, 2015)
- ❑ Evidence may be lost because the origin and authenticity of the data may be uncertain (Meghini, 2013: 51).
- ❑ Access and use restrictions (e.g. digital rights management) may not be respected in the future (Meghini, 2013: 51).

Andy. Brandon50 at https://www.flickr.com/photos/54027476@N07/4999919941/

Why should we preserve data?

Content-related reasons (Continued):

- The current custodian of the data may cease to exist at some point in the future (Meghini, 2013: 51);
- The data might have been generated using discipline specific proprietary software and hardware, that makes it impossible to read by researchers, other than the creators of the data (Meghini, 2013: 51);
- The language or knowledge of the research community may change over time, which can change or destroy the meaning of the metadata associated with the data object (Meghini, 2013: 52);
- Long-term preservation of research data saves time and finances by preventing duplication of research (Deakin University, 2017).

Why should we preserve data?

Content-related reasons (Continued)

- The meaning of a digital data object may not be evident because it does not have sufficient context information to understand its content (*The CESSDA User Guide on digital preser*vation, 2016).
- Data should be Independently Understandable: in other words it "should be is sufficiently complete to allow it to be interpreted, understood and used" by a designated research community, "without having to resort to special resources not widely available" (Giaretta, 2011: 15);
- Authenticity of a digital data object: this refers to the trustworthiness of a record, and the assurance that it is what it purports to be, and not been tampered with, or corrupted (Brown, 2013: xii)

Why should we preserve data?

Stakeholder Requirements

- Funders might have a requirement for data to be preserved long-term
- Government might require data to be preserved long-term
- Institution that generated the data might require data to be preserved (protect institution against risk of data loss, reputation)
- Contractual requirements by industry, commerce

Taking steps to Preserve Data

In the Creating Data Stage of the research data lifecycle the researcher normally compiles a Data Management Plan, that indicates what could/will be done with the data during the data life cycle. This plan will typically include a section on the preservation of the data

Taking steps to preserve data

- File formats
  - "Ensure that the file formats you use" for storage "are widely adopted", or is an open format (Deakin University, 2016)
- Migration
  - If data is stored using a format that is about to become obsolete, or might become obsolete in the future, migrate this to a more suitable format
    (See slide on File Formats)
  - Alternatively, preserve the entire environment needed to access and/or use the data, e.g. store the operating system (or virtual machine) and all required software with the data and metadata in a zip file/BagIt folder
- Software
  - Choose software that is widely used, adopted and well supported (Deakin University, 2016);
- File Storage Media
  - Ideally store your data on a network drive to ensure that data is effectively backed-up and available to be migrated to other media when needed (Deakin University, 2016)

Taking steps to preserve data

- Version control
  - Version control of a file is important to track changes to a file, especially when multiple members of a research team has access to a file.
  - If the software you use does not support version control (e.g. MS Word), then you could "set up explicit rules to ensure version tracking of files" – this could entail, keeping a master copy of the files, including date/times as part of the file names (Deakin University, 2016).
- Back-up strategies
  - If you cannot use network storage, make sure that you routinely move your data onto a fresh medium, to safeguard it against media degradation (Deakin University, 2016).
  - Make sure that you have a back-up strategy that includes multiple-site storage (replication servers), and frequently check and restore files (Deakin University, 2016)

Taking steps to preserve data

- Ownership and access
  - "Allocate responsibility for data preservation to a member" of the research team (it could be an embedded librarian) (Deakin University, 2016)
  - Determine who will need access to the "preserved data files, and who will have ongoing responsibility and ownership of the data, to avoid data loss"when the staff member moves on (Deakin University, 2016).
- File organisation and file naming conventions
  - Organise your files in a tiered folder structure – folder names and file names should clearly describe the contents
  - Save your file names according to file naming conventions (See University of Edinburgh's Naming Conventions at http://www.ed.ac.uk/records-management/records-management/staff-guidance/electronic-records/naming-conventions)

Taking steps to preserve data

- Hardware

This a very difficult decision, but some authors even suggest that it is useful to keep some of the old computer workstations, and equipment somewhere in a hardware archive/museum

Blake Patterson, at https://www.flickr.com/photos/blakespot/2498797229

Bill Bertram, at https://commons.wikimedia.org/wiki/File:C64c_system.jpg

Metadata for preservation

- Digital preservation also requires the collection, management, and use of metadata about an object that will be preserved, but the question is which metadata.
- The OAIS Reference Model is the most general conceptual framework that defines the types of metadata that are necessary for achieving digital preservation. For more information on the OAIS model go to http://bit.ly/2lxU7xj
- OAIS distinguishes between two types of metadata – **Representation Information** and **Preservation Description Information**.

Metadata for preservation

❑ Representation Information

Consist of metadata about the understandability of digital objects (Giaretta, 2011: 16-17). The representation information maps a data object into more meaningful concepts, and can further be separated into:

- **Structure Information** (the encoding format of the data object);
- **Semantic Information** (description that captures enough semantics/meaningful information about the digital object) (Meghini, 2013: 52).

Metadata for preservation

❑ Preservation Description Information

Consist of metadata about the origins, context, and restrictions of a digital object (Giaretta, 2011: 16-17). Preservation Description Information can further be divided into:

- **Reference Information**, focusing on “identification of the data object”;
- **Provenance Information**, focusing on the “history of the data object”;
- **Context Information** focusing on the “relationships of the data object with its environment”;
- **Fixity Information** focusing on “mechanisms for ensuring that the data object has not been altered”;
- **Access Right Information**, “focusing on access restrictions of the data object” (Giaretta, 2011: 21-23; Meghini, 2011: 52).

Levels of digital preservation (NDSA)
(Phillips et al., 2013)

Level 1 – Protect Your Data

Level 2 – Know Your Data

Level 3 – Monitor Your Data

Level 4 – Repair Your Data

National Digital Stewardship Alliance (NDSA)

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

Preservation file formats

- "File formats should ideally be considered and decided upon before the commencement of data collection" (ANDS Guide: file formats, 2016: 1).
- Some helpful resources that lists acceptable file formats for data preservation:

- **University of Sydney Library** table of acceptable file formats for long-term preservation, available at https://library.sydney.edu.au/research/data-management/file-formats.html
- **The UK Data Archive** file formats table - available at http://www.data-archive.ac.uk/create-manage/format/formats-table
- **National Archives of Australia** Preservation File Formats – available at http://naa.gov.au/Images/Preservation-File-Formats_tcm16-79398.pdf

University of Sydney File Formats

https://library.sydney.edu.au/research/data-management/file-formats.html

Format Category	Acceptable Formats (*preferred)
Archive	GNU ZIP File Format (.gzip); Tape Archive File Format (.tar); ZIP File Format (.zip)
Audio	Audio Interchange File Format (.aiff); *Free Lossless Audio Codec (.flac); *Waveform Audio File Format (.wav)
Computer Aided Design (CAD)	Design Web Format (.dwf); Drawing Exchange Format (.dxf); Drawing Files (.dwg, .dws, .dwt); Extensible 3D (.x3d); Standard for the Exchange of Product; Model Data (.step, .stp)
Email	*Email (Electronic Mail Format) (.eml); *MBOX Email Format (.mbox); Microsoft Outlook Item (.msg); *Microsoft Outlook Personal Folders File (.pst)
Geospatial (see also CAD and Dataset categories)	ESRI Shapefile (.shp, .shx, .dbf); Geospatial Tagged Image File Format (.tif, .tiff, .gtiff); Keyhole Markup Language (.kml)
Moving Images	Motion JPEG 2000 (.mj2); MPEG-4 (.mp4)
Presentations	Microsoft PowerPoint (.pptx); *OpenDocument Presentation (.odp)
Still Images	Portable Network Graphics (.png); *Tagged Image File Format (.tif, .tiff)
Tabular Datasets	*Comma Separated Values (.csv); eXtensible Markup Language (.xml); Microsoft Excel (.xlsx); OpenDocument Spreadsheet (.ods); Tab Delimited Values (.tab, .tsv, .txt)
Text	eXtensible Markup Language (.xml); Microsoft Word (.docx); OpenDocument Text (.odt); Plain Text (ASCII, UTF-8, or UTF-16) (.txt); Portable Document Format (.pdf); Rich Text Format (.rtf)
Website	eXtensible HyperText Markup Language (.xhtml); MIME HTML (.mhtml); Web ARChive File Format (.warc)

What should be preserved?

- Data files (preferably in open formats)
- Software programmes
- Operating systems
- Hardware
  - Desktop Computer (PC, laptop)
  - Mobile device (tablet, recorder, etc.)
  - Storage Disks (Magnetic Tapes, Cassettes, Floppy Disks, Stiffies, Flashdisks, hard drives)
  - Instruments

Institutional planning for research data preservation

- The International Federation of Data Organisations for Social Science (IFDO), lists three broad areas that institutions need to consider when addressing data preservation:
  - Organisational Infrastructure:

  "Policies, procedures, practices and people" as well as "legal and regulatory frameworks", preservation knowledge and skills, and all issues regarding "funding and resource planning"

  - Technological concerns:

  This includes equipment, software, hardware, media monitoring and refreshment strategies

  - Data curation:

  This includes pre-ingestive actions, ingest functions, archival storage and preservation, as well as the dissemination of and access to data for its designated community.

Institutional Readiness

- The National and State Libraries Australasia has developed a **Digital Preservation Environment Maturity Matrix** to assist organisations to assess their readiness or progress with regard to digital preservation. This tool is available at:

- http://www.nsla.org.au/sites/www.nsla.org.au/files/publications/NSLA.DigPres_Environment_Maturity_Matrix.pdf

- **MIT Libraries** have developed a number of **tutorials** on short term strategies for digital preservation, available at http://www.dpworkshop.org/ as well as a **survey form** that can be used to determine **institutional readiness**, available at http://www.dpworkshop.org/sites/default/files/readiness.pdf

End of Overview

Bibliography

- ***ANDS Guide: file formats.*** Caulfied-East, Melbourne, VIC: Australian National Data Service, 2016. [Online] available at http://www.ands.org.au/__data/assets/pdf_file/0003/731775/File-Formats.pdf (Accessed 12 February 2017).
- BROWN, A. 2013. ***Practical digital preservation: a how to guide for organizations of any size.*** London: Facet Publishing.
- ***The CESSDA User Guide on digital preservation***, 2016. Bergen, Norway: Consortium of European Social Science Data Archives [Online] available at http://bit.ly/2lxU7xj (Accessed 12 February 2017).
- CHOUDHURY, S. 2014. ***Public Institution perspective (Research Library).*** Presented at the Digital Media Analysis, Search and Management (DMASM), 2014. [Online] available at http://dataconservancy.org/wp-content/uploads/2014/03/DC_DMASM_2014.pdf (Accessed 24 September 2014).

Bibliography

- ***Data preservation***, n.d. Caulfied-East, Melbourne, VIC: Australian National Data Service. [Online] available at http://www.ands.org.au/working-with-data/data-management/data-preservation
- Deakin University. 2016. ***Preserve.*** Geelong, Victoria, Australia: Deakin University. [Online] available at http://www.deakin.edu.au/students/research/research-support-and-scholarships/eresearch/manage-data/preserve (Accessed on 12 February 2017).
- GIARETTA, D.L. 2011. ***Advanced digital preservation.*** Berlin, Germany: Springer
- ***International Federation of Data Organizations for Social Scien***ce, 2012. [Sl.: ns] [Online] available at http://ifdo.org/wordpress/?page_id-18 (Accessed 12 February 2017).

Bibliography

- LEBOW, M. AND CARUSO, M. 2015. ***Digital preservation and data repositories: just what does "long-term" mean anyway?***. Presentation given at the 2015 DLF Forum in Vancouver, BC in snapshot session. Seattle, WN: Washington University. [Online] available at http://hdl.handle.net/1773/34272 (Accessed 12 February 2017)
- National Digital Stewardship Alliance. n.d. ***Levels of digital preservation.*** Washington, DC: NDSA c/o CLIR+DLF. [Online] available at http://ndsa.org/activities/levels-of-digital-preservation/ (Accessed 12 February 2017).
- Phillips, M. et al. 2013. The NDSA levels of digital preservation: an explanation and uses. ***Proceedings of the Archiving (IS&T) Conference,*** April 2013, Washington, DC. [Online] available at http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf (Accessed 12 February 2017).