THE GENERATION OF KNOWLEDGE · TRUST, QUALITY AND THE FUTURE

# Preserving Knowledge for the Future: Archives and Digital Preservation

Why knowledge is lost, how archival thinking developed, why digital information is more fragile than it seems, the models and practices used to preserve it, the evidence on disappearing data and links, and how research organisations can keep their data and findings usable for decades.

ModQu research article · 2,515 words · 24 references (APA 7th edition)

Knowledge that is not preserved is eventually lost. The library of Alexandria is remembered because of what disappeared with it, but most losses are quieter: records discarded because no one saw their value, files left on obsolete media, datasets that no one can now interpret. Each generation of knowledge depends on what earlier generations kept. For research, preservation matters for another reason as well. Findings that cannot be checked against the data and methods that produced them cannot be fully trusted, and data that cannot be found again cannot be reused to answer new questions. This article examines how thinking about archives developed, why digital information is especially at risk, and the models and practices that help keep knowledge available for the future.

## Archives and their purposes

Archives have long served two purposes: to keep evidence of what was done and decided, and to preserve a record of the past for later generations. Jenkinson (1922), in an influential manual written in England, emphasised the first. He saw the archivist as the impartial custodian of records created in the course of business, whose duty was to keep them intact, in their original order and free from interference, so that they could serve as trustworthy evidence. The archivist, in his view, should not choose which records to keep, since that would introduce bias.

Schellenberg (1956), writing in the United States after the vast growth of government records in the twentieth century, argued that archivists could not keep everything and must therefore **appraise** records to decide which to preserve. He distinguished between the primary value of records to the organisation that created them and their secondary value as evidence of how the organisation functioned and as information about people, places and events for researchers. Appraisal became a central archival task.

Cook (1997), a Canadian archivist, traced the history of archival ideas over the twentieth century and argued that the profession was moving from a focus on records as evidence for their creators towards a broader view of archives as society's memory. In this view, archivists make active choices about what the future will be able to know, and they bear responsibility for ensuring that archives reflect the whole of society rather than only its most powerful institutions.

## Why digital information is fragile

Digital technology seemed at first to solve the problem of preservation. Copies are perfect, storage is cheap and information can be duplicated without limit. Experience showed otherwise. Rothenberg (1995) observed wryly that digital information lasts forever, or five years, whichever comes first. He identified the core problem: digital documents depend on the hardware and software needed to read them. Storage media decay or become obsolete, file formats fall out of use, and programs that once interpreted the data cease to run on new computers. Unlike a paper document, which can be read as long as it survives, a digital file may survive physically and yet be impossible to read.

Hedstrom (1997) described digital preservation as a "time bomb" for digital libraries. She noted that the volume of digital material was growing far faster than the capacity of institutions to manage it, and that the methods available for preserving it, such as migrating files to new formats or emulating old systems, were costly and untested at scale.

A report by Waters and Garrett (1996), prepared for the Commission on Preservation and Access and the Research Libraries Group in the United States, defined the problem in terms that still guide the field. They argued that the responsibility for preserving digital information must be shared between those who create it and those who archive it, that preservation requires a deliberate and continuing effort rather than a single act, and that a network of trusted digital archives was needed. They emphasised that preserving digital information means preserving not only the bits but the context needed to interpret them: who created the material, how and why.

In 2015, the computer scientist Vint Cerf warned that the twenty-first century risked becoming a **digital dark age**, in which future historians would know less about our time than about earlier centuries, because so much information would have become unreadable (Ghosh, 2015).

## The evidence of loss

Studies of scholarly communication show how quickly digital knowledge disappears. Vines et al. (2014) tried to obtain the data underlying 516 studies in biology published between 2 and 22 years earlier. The availability of data fell rapidly with the age of the article: the odds that the data still existed fell by about 17% for each year since publication. Often the data were simply lost, stored on old computers or media that no longer worked. The odds of reaching an author through the email address in the article also declined each year.

Links to web resources decay as well. Klein et al. (2014) examined millions of web links cited in scholarly articles and found that one in five articles suffered from **reference rot**: at least one of the web resources they cited had disappeared or changed so that it no longer contained the content originally referred to. The problem grew worse with the age of the article. Because scholarship depends on the ability to follow references back to their sources, reference rot weakens the chain of evidence on which knowledge rests.

## Preserving the web

Much of today's public record exists only on the web, where pages are changed or removed every day. Kahle (1997), who founded the Internet Archive in 1996, argued that the web was becoming a central part of the cultural record and that, unless it was systematically collected, future generations would have no record of it. The Internet Archive began to

save copies of publicly available web pages, and its collection later became accessible to the public through the Wayback Machine. National libraries in many countries now collect websites under their legal deposit responsibilities. Masanès (2006) described the methods involved, from automated crawling of large parts of the web to selective capture of particular sites, and the difficulties they face: dynamic pages that change with every visit, content hidden behind logins or search forms, and the sheer scale of material. Web archives cannot capture everything, but they preserve a great deal that would otherwise vanish, including many of the sources that scholarly articles cite.

## The OAIS reference model

The most influential framework for digital preservation is the **Open Archival Information System** (OAIS) reference model, developed by space agencies through the Consultative Committee for Space Data Systems and adopted as an international standard (CCSDS, 2012). The model describes what an archive must do to preserve information for a designated community of users over the long term.

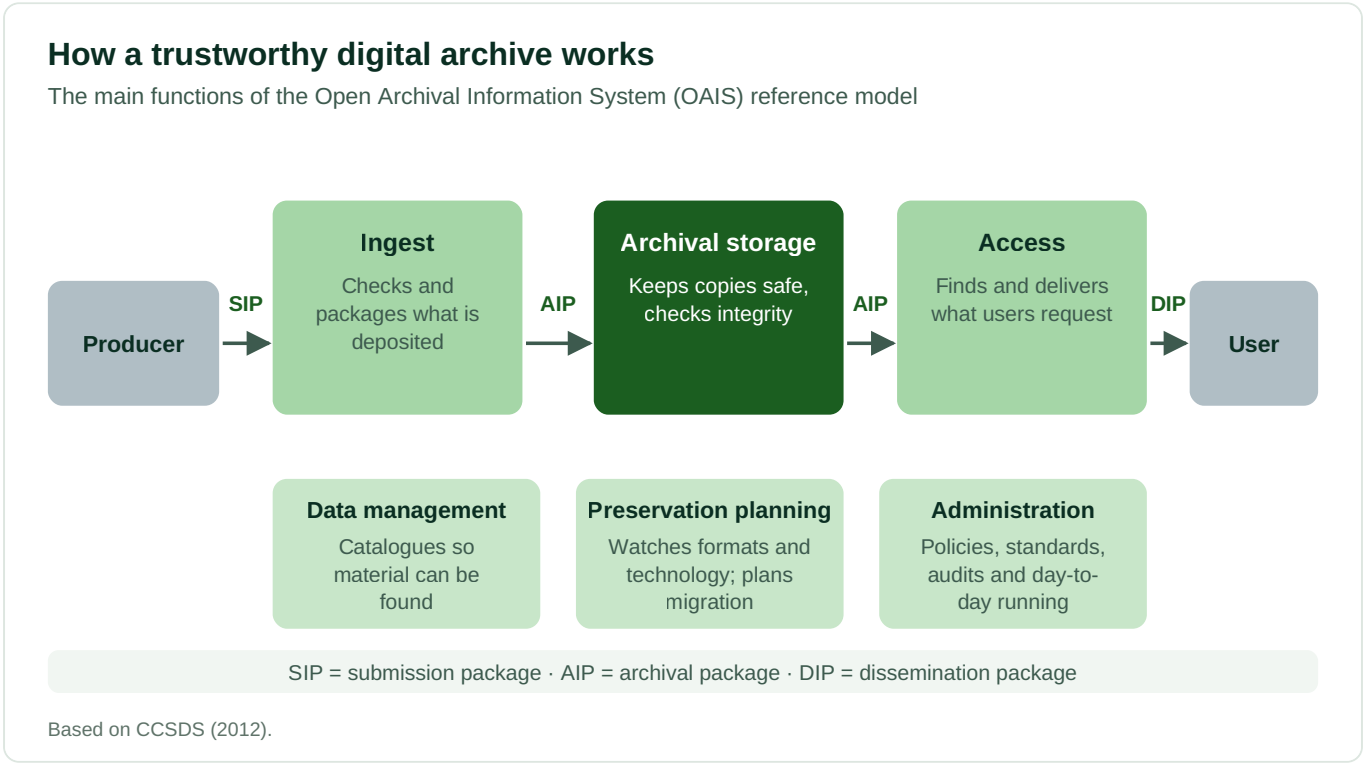



Figure 1. The main functions of the OAIS reference model for digital archives. Based on CCSDS (2012); diagram by ModQu.

In the OAIS model, information arrives from a producer as a **submission information package**. The archive checks it and converts it into an **archival information package**, which contains both the content and the information needed to understand and preserve it, such as descriptions of its format, its origin and how it has been changed. When a user requests the material, the archive delivers a **dissemination information package** suited to the user's needs. Supporting functions manage the catalogue, plan for changes in technology and administer the archive. Lavoie (2004) explained that

one of the model's most important contributions was its insistence that an archive must preserve not only data but the **representation information** that makes the data understandable to its intended users, and that it must monitor its community and technology to make sure that this remains so.

Vardigan and Whiteman (2007) described how the Inter-university Consortium for Political and Social Research (ICPSR), one of the oldest social science data archives, applied the OAIS model. They found that the model fitted the archive's practices well and helped to identify gaps, for example in documenting the steps taken to process and preserve each dataset.

## Practices of digital preservation

Several practices have become standard.

**Multiple copies.** Keeping several copies in different places protects against loss through hardware failure, accident or disaster. Reich and Rosenthal (2001) described LOCKSS ("Lots of Copies Keep Stuff Safe"), a system in which libraries each hold copies of the same journals and regularly compare them, repairing any copy that has been damaged.

**Integrity checking.** Archives compute digital fingerprints, known as checksums, for each file and check them regularly to detect corruption. Rosenthal et al. (2005) argued that digital preservation systems should be designed on the assumption that every component, including storage, software and staff, will eventually fail, and that the system must detect and recover from such failures.

**Format management.** Archives prefer open, well-documented file formats that are likely to remain readable, and they plan to migrate files to new formats when old ones become obsolete. For some material, they preserve the original software and use emulation to run it on new hardware.

**Documentation.** Data are only useful if they can be understood. Archives therefore preserve descriptions of how data were collected, what each variable means and how files are organised. The curation lifecycle model of the Digital Curation Centre describes preservation as a continuing cycle of creation, appraisal, ingestion, storage, access, reuse and transformation, with description and planning running throughout (Higgins, 2008).

**Persistent identifiers.** Assigning permanent identifiers, such as digital object identifiers, to datasets and publications allows them to be found even when their web addresses change, reducing the problem of reference rot.

**Certification.** Users need assurance that an archive is trustworthy. Standards such as ISO 16363 set out criteria against which repositories can be audited, covering their organisation, their management of digital objects and their technical infrastructure (International Organization for Standardization, 2012).

## Deciding what to keep

Not everything can or should be kept. Schellenberg's principle of appraisal applies to digital material as much as to paper. Whyte and Wilson (2010), in guidance for the Digital Curation Centre in the United Kingdom, proposed criteria for deciding which research data to preserve. They include the relevance of the data to the mission of the organisation that would hold them, their scientific or historical value, whether they are unique or could be recreated, whether they can

be reused given the documentation available, whether there are legal or ethical obligations to keep or destroy them, and the cost of preserving them. Data from a one-off observation, such as a survey of public attitudes at a particular moment, are usually worth keeping because they cannot be gathered again. Intermediate files that can be regenerated from the raw data and the analysis code are often not.

Appraisal also means deciding what to discard. Keeping everything has costs: storage, management and the difficulty of finding what matters among what does not. Personal data that are no longer needed may also create risks for the people they describe, and data protection law in many countries requires that such data be kept no longer than necessary. A good preservation policy therefore states what will be kept, for how long and why, and what will be deleted.

## The economics of preservation

Preservation costs money, and the costs continue indefinitely. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010) argued that the central problem of digital preservation is economic rather than technical. The benefits of preservation often accrue to future users who do not pay for it, while the costs fall on present organisations whose priorities lie elsewhere. The task force argued that sustainable preservation requires clear decisions about what is worth preserving, clear responsibility for preserving it, and funding models that link the costs to those who benefit.

Lynch (2003) argued that universities should take responsibility for preserving the research their members produce, by building **institutional repositories** that collect, manage and preserve digital materials. Such repositories have become common, although many struggle to secure the long-term commitment that preservation requires.

## Authenticity and trust

Preserving knowledge is not only a matter of keeping data readable. Users must be able to trust that what they find is what was originally created and that it has not been altered. Duranti and Rogers (2012) examined how records held in cloud services raise new questions about authenticity and legal accountability, because organisations may not know where their records are stored, who can access them or how they are being maintained. Archives address these concerns by documenting the custody of records, recording every change made during preservation and using checksums and audit trails to show that content has not been altered.

## Preserving research data

For survey and social research, preservation has particular importance. Survey data are expensive to collect and cannot be collected again for the same time and place. A survey of public opinion in 2024 records something that will never be observable again. Preserved data allow later researchers to check original findings, to combine data across time and countries, and to ask questions the original researchers did not think of.

Corti et al. (2019), in a guide to managing and sharing research data, emphasised that preservation begins at the start of a project, not at its end. Researchers should plan how data will be documented, formatted and stored; obtain consent from participants that allows for archiving and sharing where appropriate; remove information that could identify individuals;

and deposit data with a trusted archive along with the questionnaires, codebooks and methodological reports needed to understand them. The FAIR principles, which call for data to be findable, accessible, interoperable and reusable, provide a widely accepted set of goals for this work (Wilkinson et al., 2016).

## Implications for research organisations

For organisations that generate knowledge through surveys and research, several practices follow:

- **Plan for preservation from the start.** Decide what will be kept, in what formats and for how long, before data are collected.
- **Document thoroughly.** Keep questionnaires, sampling details, weighting procedures and variable definitions with the data, so that future users can understand them.
- **Use open formats and multiple copies.** Store data in widely supported formats and keep copies in more than one place, with regular integrity checks.
- **Assign persistent identifiers.** Make datasets and reports citable with permanent identifiers so that references to them do not break.
- **Deposit with trusted archives.** Where possible, place data and documentation with certified repositories that have a long-term mission.
- **Respect confidentiality.** Remove identifying information and follow the terms of participants' consent when preserving and sharing data.

## Conclusion

Preserving knowledge is an active, continuing task. Archival thinking moved from the custody of evidence, through the appraisal of records, to a broader view of archives as society's memory. Digital technology made copying easy but preservation hard, because digital information depends on hardware, software and context that change quickly. Research shows that data and links disappear rapidly unless they are deliberately preserved. Models such as OAIS, and practices such as multiple copies, integrity checking, format management, documentation, persistent identifiers and certification, provide a basis for keeping knowledge usable. The remaining challenges are largely economic and organisational. For producers of research, preservation is part of producing knowledge well: findings that can be checked, and data that can be reused, are the foundations on which future knowledge will be built.

## References

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable economics for a digital planet: Ensuring long-term access to digital information*. Blue Ribbon Task Force.

CCSDS. (2012). *Reference model for an open archival information system (OAIS)* (CCSDS 650.0-M-2, Magenta Book). Consultative Committee for Space Data Systems.

Cook, T. (1997). What is past is prologue: A history of archival ideas since 1898, and the future paradigm shift. *Archivaria, 43*, 17–63.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice* (2nd ed.). SAGE.

Duranti, L., & Rogers, C. (2012). Trust in digital records: An increasingly cloudy legal area. *Computer Law & Security Review, 28*(5), 522–531. https://doi.org/10.1016/j.clsr.2012.07.009

Ghosh, P. (2015, February 13). Google's Vint Cerf warns of "digital Dark Age". *BBC News*.

Hedstrom, M. (1997). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities, 31*(3), 189–202. https://doi.org/10.1023/A:1000676723815

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

International Organization for Standardization. (2012). *Space data and information transfer systems: Audit and certification of trustworthy digital repositories* (ISO Standard No. 16363:2012). ISO.

Jenkinson, H. (1922). *A manual of archive administration*. Clarendon Press.

Kahle, B. (1997). Preserving the Internet. *Scientific American, 276*(3), 82–83. https://doi.org/10.1038/scientificamerican0397-82

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLoS ONE, 9*(12), e115253. https://doi.org/10.1371/journal.pone.0115253

Lavoie, B. F. (2004). *The Open Archival Information System reference model: Introductory guide* (DPC Technology Watch Report 04-01). Digital Preservation Coalition.

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *portal: Libraries and the Academy, 3*(2), 327–336. https://doi.org/10.1353/pla.2003.0039

Masanès, J. (Ed.). (2006). *Web archiving*. Springer. https://doi.org/10.1007/978-3-540-46332-0

Reich, V., & Rosenthal, D. S. H. (2001). LOCKSS: A permanent web publishing and access system. *D-Lib Magazine, 7*(6). https://doi.org/10.1045/june2001-reich

Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine, 11*(11). https://doi.org/10.1045/november2005-rosenthal

Rothenberg, J. (1995). Ensuring the longevity of digital documents. *Scientific American, 272*(1), 42–47. https://doi.org/10.1038/scientificamerican0195-42

Schellenberg, T. R. (1956). *Modern archives: Principles and techniques*. University of Chicago Press.

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: Applying the OAIS reference model to the social science archive context. *Archival Science, 7*(1), 73–87. https://doi.org/10.1007/s10502-006-9037-z

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S.,
Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Waters, D., & Garrett, J. (1996). *Preserving digital information: Report of the Task Force on Archiving of Digital Information*. Commission on Preservation and Access.

Whyte, A., & Wilson, A. (2010). *How to appraise and select research data for curation* (DCC How-to Guides). Digital Curation Centre.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18