ESSENTIAL DIGITAL PRESERVATION

Part II

FUNDAMENTAL IDEAS ABOUT PRESERVING DIGITALLY ENCODED INFORMATION

DAVID GIARETTA

david@giaretta.org

www.giaretta.org and www.iso16363.org

This book is being published in 5 sections.

Part 1 : Why read this book? Why is preserving digitally encoded information important - but difficult.

Part II : Fundamental ideas about preserving digitally encoded information

Part III: What to do and when to do it, to preserve digitally encoded information

Part IV: Adding Value and Exploiting Information

Part V: Evaluating claims about preserving digitally encoded information

Contents

1 Thinking clearly about what is needed for digital preservation 5

1.1 What digital preservation must address 5

1.2 Various suggested ways to preserve digitally encoded information 6

1.3 What can we rely on? 7

1.4 What people rely on currently, but may change in the future 8

1.4.1 Software, and its configuration 8

1.4.2 Hardware 9

1.4.3 Tacit knowledge 11

1.4.4 Remote resources 11

1.4.5 Slow changes 12

1.5 Common errors in digital preservation 12

1.6 Limitations of this chapter 13

2 What OAIS says and why it says it 14

2.1 The aims of OAIS 14

2.2 Introduction to OAIS concepts and terminology 15

2.3 Preserve what, for how long and for whom? 15

2.4 OAIS Information Model component definitions 16

2.4.1 What *"metadata"*, how much *"metadata"*? 20

2.4.2 Recursion – a pervasive concept 22

2.4.3 Preservation Description Information 25

2.4.4 Information Packaging 32

2.5 OAIS Functional Model 35

2.6 Limitations of this chapter 40

3 Understanding what OAIS conformance actually means 41

3.1 What OAIS says about conformance 41

3.2 OAIS Information Model 41

3.3 OAIS Mandatory Responsibilities 42

3.3.1 Negotiate for and accept appropriate information from information Producers. 42

3.3.2 Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation. 42
3.3.3 Determine, either by itself or in conjunction with other parties, which entities should become the Designated Community. i.e., the communities that should be able to understand the information provided. Definition of the Designated Community includes a determination of their Knowledge Base. 43
3.3.4 Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information. 43
3.3.5 Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions. 44
3.3.6 Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Content Information with evidence supporting its Authenticity. 44
3.4 Limitations of this chapter 45
4 Explanation of updates to OAIS version 2012 46
4.1 Updates to OAIS Concepts 46
4.1.1 Representation Information 46
4.1.2 Preservation Description Information (PDI) 46
4.1.3 Preservation Objectives 47
4.2 Update to the OAIS Functional Model 48
4.3 Updates to the OAIS Information Model 48
4.3.1 Updates to Information Package Definition 49
4.4 Updates to Preservation Perspectives 50
4.5 Updates to Archive Interoperability 50
4.6 Limitations of this chapter 51
5 Authenticity 52
5.1 What is authenticity? 52
5.2 Background to Authenticity 52
5.2.1 Links to previous literature 53
5.3 OAIS Definition of authenticity 54
5.3.1 Transformational Information Properties 56
5.4 Types of evidence to support claims of Authenticity 57
5.5 Limitations of this chapter 58
6 Lifecycles 59
6.1 So many Lifecycle models 59
6.2 Limitations of this chapter 61

7 Active Data Management Plans, collecting Information needed for Re-Use and Preservation – what to capture and when ... 62
7.1 (Active) Data Management Plans ... 62
7.1.1 Monitoring the ADMP ... 64
7.1.2 Information from planning stages of the project ... 65
7.1.3 Creation ... 66
7.1.4 Primary use ... 67
7.1.5 Handover (to a Repository) ... 67
7.1.6 Re-use ... 67
7.1.7 Preservation ... 68
7.2 Information Preparation to Enable Long Term Usability (z) ... 68
7.2.1 Initiating Collection Group ... 68
7.2.2 Planning Collection Group ... 69
7.2.3 Executing Collection Group ... 69
7.2.4 Closing Collection Group ... 69
7.2.5 Control Collection Group ... 70
7.2.6 Provenance Information ... 76
7.2.7 Context Information ... 77
7.2.8 Reference Information ... 77
7.2.9 Descriptive Information ... 77
7.2.10 Packaging Information ... 77
7.2.11 Supplementary Information held by the archive ... 78
7.3 Limitations of this chapter ... 78

PART II: FUNDAMENTAL IDEAS ABOUT PRESERVING DIGITALLY ENCODED INFORMATION

1 Thinking clearly about what is needed for digital preservation

A fundamental question is what we mean by digital preservation? This chapter describes what digital preservation needs to address.

1.1 What digital preservation must address

The definition of digital preservation proposed by UNESCO[1] is

> *Digital preservation consists of the processes aimed at ensuring the continued accessibility of digital materials. To do this involves finding ways to re-present what was originally presented to users by a combination of software and hardware tools acting on data.*

There are many, similar definitions which talk about accessibility but have slightly different emphases.

What many seem to mean is that one can do the same as with paper, namely in 50 years one can print the digital document and look at the symbols on the page. That may be acceptable for typical documents in libraries. We also need to be sure the symbols on the page have not been altered. In the physical world we can examine the medium. For example, sheepskin was chosen as the medium for legal documents in previous centuries because its fat content and the tendency of sheepskin to delaminate[2] meant that "*they do not easily yield to erasure without the blemish becoming apparent*."[3]

One also would also like to be sure that the paper has not been replaced; one way to do this would be to give it to a trusted person or organisation, such as a lawyer or bank, to keep in a vault with restricted access. One would like similar assurances in the case of a digitally encoded document, in terms of being sure that the digital object is what we think it is and has not been changed. In a library, a paper document is catalogued and information such as the language in which it is written is recorded. Once the printing press was invented it became possible to print hundreds, or thousands, and even millions, of the same page, so if one copy was lost then we could be sure there were many others.

Of course, in the past, before the digital world had been invented, there was no alternative to print and so everything could be preserved in the same way as the document mentioned at the start.

But we are thinking now about the digital world where the amount of information is so large that it would not be possible, nor would it be useful, to print it **all**.

In the case of the documents, it seems reasonable be satisfied if we are able, in the future, to do what we can do now. Of course, in the case of a digitally encoded document one would need software to be able to display and print the contents. That software will need to work

[1] See https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital-preservation

[2] See https://theconversation.com/sheepskin-was-used-as-an-anti-fraud-device-in-british-legal-documents-for-hundreds-of-years-158547

[3] Dialogus de Scaccario, The Dialogue concerning the Exchequer Late 12th Century, see https://sourcebooks.fordham.edu/source/excheq1.asp

into the future somehow. Also, the provenance of the document these days may be encoded in some way, for example using PREMIS[4], and that will also need to continue to be understood, including any special vocabulary. In other words, whereas when everything was on paper (or the equivalent) then one could simply say that it is up to a human to read it, but when things are digital “no document is an island”, to coin a phrase.

What about things that are not normally printed?

As an example, let us consider all the astronomical data produced by an astronomical observatory, which would currently be in the form of FITS files. We can certainly do the equivalent of printing it and looking at it, even if we only printed the “1”s and “0”s, but is that enough?

Thinking about the astronomical data in FITS file, what would someone who is familiar with it normally do? The answer is that the data would be analysed, perhaps in combination with other data. If we want to do the same thing in 50 years’ time, what do we need to do?

The obvious thing to do is to try to make sure that the software which is used currently is available in the future. That way people who are familiar with the astronomical data will be able to use the software to process the data in the way they currently do.

But what about astronomers who are not familiar with that software? They could write their own software – but they would have to know a lot about the data, especially the details of the format and also where to find information such as where the astronomical instrument was pointing, what wavelengths were being observed, what the resolution of the image is, where the data was collected, what processing has the data been through already, and so forth – much of which may be, in fact are, encoded in the data file, as shown in kater sections.

1.2 Various suggested ways to preserve digitally encoded information

Some preservation systems claim that they can preserve by keeping a great many copies, however this is simply keeping the bits but not the software or semantics which are needed for usability.

Some people, particularly software vendors, say that we can rely on software vendors to take care of preservation but software by itself needs to be configured and managed.

Some hope there a ‘magic’ preservation form e.g., XML, however this is just a format, in the case of XML the tags require semantics to be associated with them, and the semantics and structure is tied into software which may be used.

Some propose that we can rely on some special type of media which will last for a hundred or even a thousand years. Vendors have proposed many “thousand year” media. The problem has been that one needs hardware, in the form of a reader, and associated software, to extract the data from the media. As an example, the BBC Doomsday Project[5] produced two special videodisc which could be read using a LV-ROM player. However, after 16 years of use most LV-ROM players had reached the end of their working lives and as a storage technology LV-ROM had been superseded by CD-ROM and DVD, leaving the Domesday discs essentially unreadable. Some long-lived media can be read with the human eye, including writing 2 points (~0.7mm) characters on Silicon Carbide sheets which are readable by the naked eye or using a magnifying glass[6], or writing small characters or even “1” and “0” on microfilm[7].

[4] See https://www.loc.gov/standards/premis/

[5] See https://discovery.nationalarchives.gov.uk/details/r/C16160

[6] Aoki K., et al., 2008, Durable Media for Long-Term Preservation of Geological Repository Records – 8320, WM2008 Conference, available at https://archivedproceedings.econference.io/wmsym/2008/pdfs/8320.pdf

[7] See https://www.piql.com/

Both have relatively low information densities[8] but most importantly they lack the semantics and software which are necessary to use the information, beyond what could be undertaken by a human reader.

Since the longevity of hardware readers and associated software have been mentioned as issues, some have suggested saving obsolete hardware and software, however keeping hardware alive does not seem like a very long-term solution since integrated circuit components are difficult to salvage or make anew.

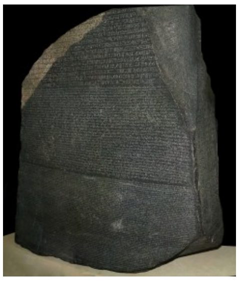
A mantra in the library community used to be "emulate or migrate[9]", which meant, if one had a data file that required some specific software then over time the underlying system (computer CPU, libraries, operating system) on which the software would become unavailable and so the alternatives would be to "emulate" the underlying system – thereby be able to continue to use the software, and if/when that is not practical then one could "migrate" the data file to a form which could be used in contemporaneous software. This approach has worked well in the library community, dealing with documents in various word processing systems, but does not work well with, for example, scientific data, because there is a loss of information with transformations, particularly semantics, and a more detailed strategy is needed.

1.3 What can we rely on?

A number of things have a track record of being usable over decades or centuries, or even millennia.

Words or symbols carved in rock that people can "read", such as the Rosetta Stone, have a record of preserving written information for thousands of years. Vellum and writing on sheepskin have lasted for many hundreds of years.

Carvings in stone and books have proven track records of preserving information over hundreds of years.



and

However, the "information density" is not great and this would not be suitable for huge amounts of information.

We hope that computers of some kind, probably very different from those used today, will be available.

We can probably assume that there will be remote access using some kind of networks. However, the networks may not use TCP/IP v4 or even v6, the identifiers we use currently may not all resolve as they do today.

We should also be able to assume there will be people, but perhaps speaking different languages than your text is written in, with very different background knowledge. In the same

[8] https://www.piql.com/about/technology/ (downloaded at 10 Sept 2022) states that "binary code is converted to grey pixels that are then written to film in four levels of grey, with 8 million data points per frame. Each film has capacity for 120 GB."

[9] S Granger, Emulation as a Digital Preservation Strategy, 2000, http://www.dlib.org/dlib/october00/granger/10granger.html

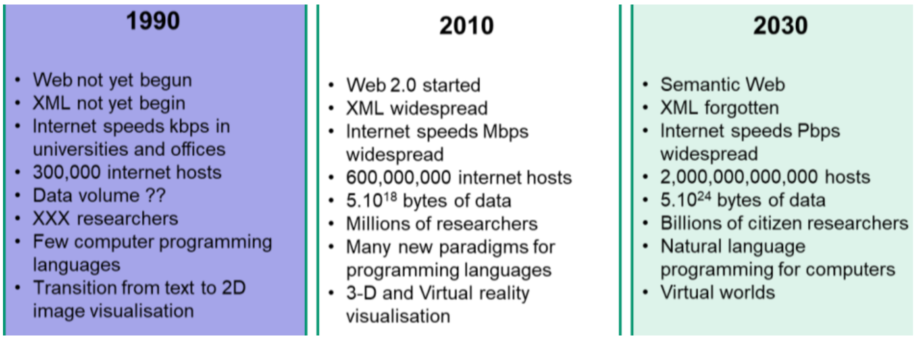
way there will also be organisations, but with different names, structures, and relationships than currently exist.

1.4 What people rely on currently, but may change in the future

Complementing section 1.3, we turn now to things which are relied upon. These things are often ubiquitous and are thereby forgotten or at least never thought about.

For all these things it is important to ask oneself – was this available to be relied on 10 or 20 years ago? – and is it certain that these can be relied on 10 to 20 years into the future?

To get some perspective, especially for younger readers, as well as readers in the future, consider the following diagram which shows some historical information plus some projected numbers, in 20-year intervals:



1990

- Web not yet begun
- XML not yet begin
- Internet speeds kbps in universities and offices
- 300,000 internet hosts
- Data volume ??
- XXX researchers
- Few computer programming languages
- Transition from text to 2D image visualisation

2010

- Web 2.0 started
- XML widespread
- Internet speeds Mbps widespread
- 600,000,000 internet hosts
- 5.10^{18} bytes of data
- Millions of researchers
- Many new paradigms for programming languages
- 3-D and Virtual reality visualisation

2030

- Semantic Web
- XML forgotten
- Internet speeds Pbps widespread
- 2,000,000,000,000 hosts
- 5.10^{24} bytes of data
- Billions of citizen researchers
- Natural language programming for computers
- Virtual worlds

Figure 1 20-year snapshots and predictions

There have been many amazing changes from 1990, and we may imagine, and in many ways expect even more changes into the future. The following sections explore some of these potential changes.

1.4.1 Software, and its configuration

We all rely on software to process or display aspects of digital objects, although as shown in the previous chapter some things can be done manually in simple cases, e.g., if we know the encoding is ASCII-7, although in special cases heroic human computational can succeed.[10]

Web browsers are commonly used, and they automatically display documents, whether HTML or Word or PDF. Some web browsers have applications built-in, but it is possible to configure the browser to open whatever application one wants for a specific content type. However, mistakes can be made and what is shown on the display by a particular browser is incorrect[11]. This can happen because a web browser sometimes has to guess how to display what it has been sent and may guess wrong[12].

A particular piece of software is not something that works in isolation. It relies on many, many, software libraries, often written and supported by others, all of which rely on an operating system. If the software is licensed, provided, and supported by a commercial company, such a company may cease to trade, which may happen for financial or legal reasons. In this case the software may cease to operate immediately, if the license is checked every time the application runs; alternatively, the software will eventually stop being usable

[10] The Human Computers of Los Alamos, see https://www.atomicheritage.org/history/human-computers-los-alamos

[11] https://www.smashingmagazine.com/2012/06/all-about-unicode-utf8-character-sets/

[12] https://deliciousbrains.com/how-unicode-works/

because updates, for example to fix security issues or to adapt to changes in software libraries on which it depends. Apart from the licensing issue, the same applies to open-source software, although in that case support for the software is often provided by individuals on a voluntary basis and those individuals may decide at any point to move their efforts to a more interesting project.

Software can depend on remote resources, and if those resources are not available, perhaps because the resource (e.g., a file) has been moved or deleted, then the software may suddenly, or perhaps a little later if it has been cached, stop working[13].

1.4.2 Hardware

The software mentioned above run on a variety of computers ranging from hand-held to desktop to remote devices. Digital hardware has changed over the past 30+ years, from CPUs and speed to devices to memory capacities to storage to network speeds. The hardware has enabled great changes in the way computers are used and how widely distributed they are. Increasing capacity has been accompanied by dramatically decreasing cost:

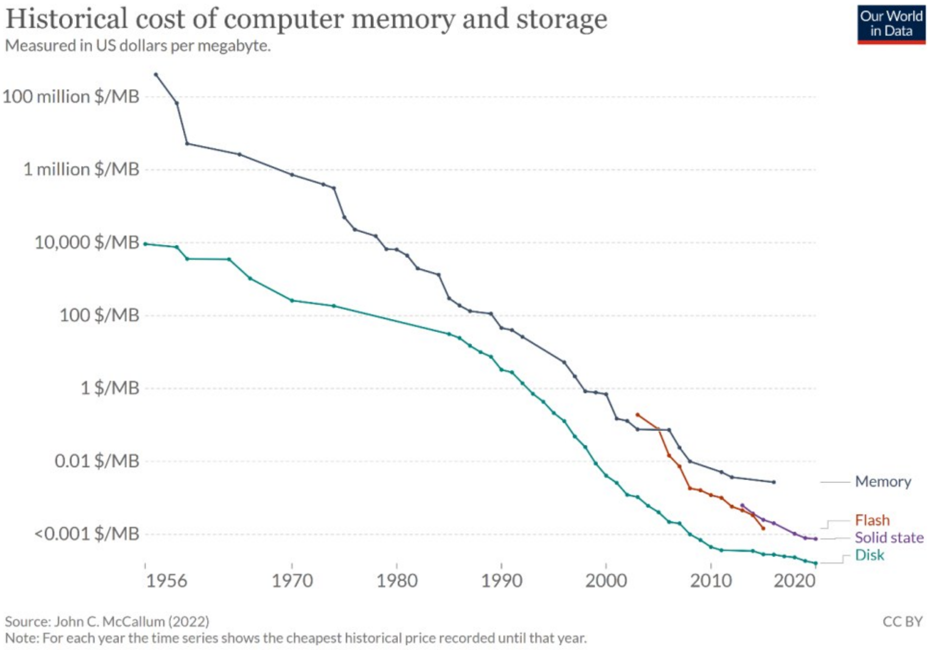


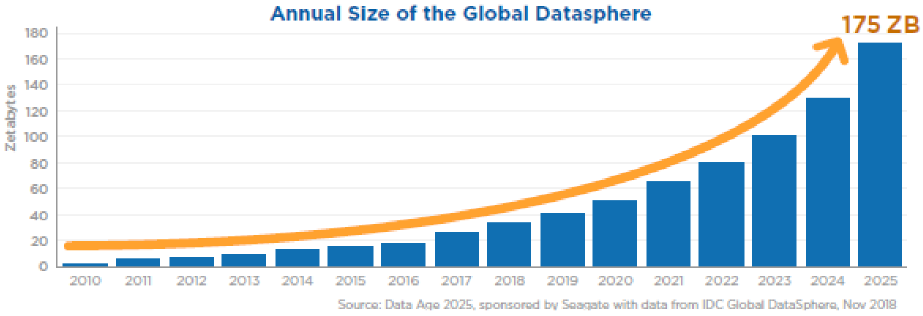
Figure 2 Cost of computer memory and storage over time - log scale

In the early 1940s, IBM's president, Thomas J Watson, reputedly said: "I think there is a world market for about five computers[14]. Now there is an expectation that almost all humans will possess, indeed hold in their hand, a computer in the form of a smart phone, which is many times more powerful than the computers that Watson was talking about.

[13] To DTD or not to DTD, 2007, post by Chris Finke, available from http://web.archive.org/web/20080206214640/http://blog.netscape.com/2007/01/16/to-dtd-or-not-to-dtd/

[14] https://www.theguardian.com/technology/2008/feb/21/computing.supercomputers

The growth of digital data has been explosive[15]:



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Figure 3 Growth of amount of data

In the same way network speed has increased – Nielson's Law suggest the growth is 50% per year[16].

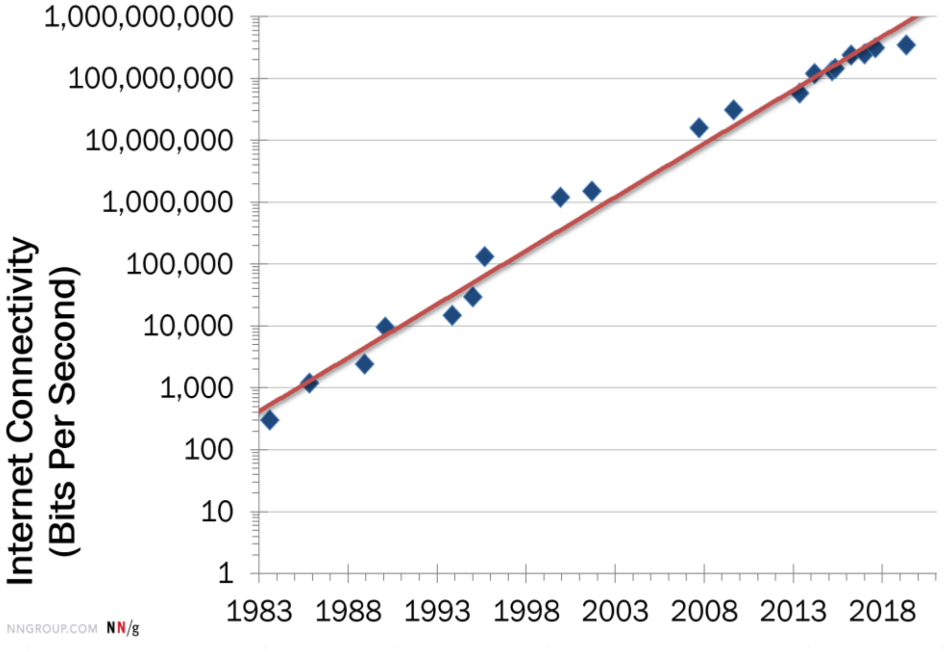


Figure 4 Growth in network speed

New operating systems have evolved, and have been created, to take advantage of the new hardware capabilities. Operating system capabilities in turn affect software libraries and applications.

The growth of data, CPU power, network speed and connectivity has enabled the growth of Cloud computing and storage, and the provision of services delivered through the cloud.

[15] From Reinsel D., Gantz J., Rydning J., Data Age 2025, 2018, available from https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[16] Nielsen J, Nielsen's Law of Internet Bandwidth, 2019, available from https://www.nngroup.com/articles/law-of-bandwidth/

1.4.3 Tacit knowledge

By tacit knowledge[17] we mean things we (i.e., a specific group, culture, nationality, age-group etc.) “all know”, ranging from concepts, the meaning of words, the meaning of slang or specialised terms or how to do certain things. Tacit knowledge tends to be viewed as intuitive knowledge which is rooted in experience, context, and practice. It tends to be hard to communicate because it resides in the mind of one of a group of individuals.

It is clear that tacit knowledge is different between different groups. It also changes over time within a group, as shown if one thinks about what one’s grandparents’ tacit knowledge must have included, and how different it would be from one’s own. In my case my grandparents would have known how to speak Italian and German, knew how to make their way around Venice, recognize their friends and family members, and do so on – none of which is known to me. Similarly, the tacit knowledge of a group of astronomers will have evolved over time as the common ideas and theories will have changed over time.

In order to be shared, such tacit knowledge must be codified and made explicit so that it can be shared.

1.4.4 Remote resources

Many of the resources one relies on in the modern world, such as those on the internet referred to by software or within data formats or as documentation, may become inaccessible or cease to exist.

An interesting example the websites of the EU projects which contributed to this book were ones for which the author owned the domain name and which he maintained, are no longer available. The reason that they are no longer available is informative. It is that the author lives in the UK and following the UK’s departure from the EU, he was no longer allowed to own “.eu” domain names[18]. This made the original websites inaccessible. Some of the project domain names, such as www.prelida.eu and www.scidip-es.eu, have been bought by others and are being used for their own purposes, unrelated to the original projects. The same would have happened if/when the author stopped paying for the ownership of the domain, or the webserver. The same could happen for any website or any remote resource. Various studies[19], [20] have shown that the half-life of URLs referenced in scientific papers is measured in years rather than decades.

It may be hoped that persistent identifiers are a solution to these issues, but all require money and effort to maintenance the pointers to the location of the remote resource. For example, the various persistent identifiers based on the handle system[21],[22], such as the Digital

[17] The Cambridge Dictionary defines “Tacit Knowledge” as knowledge that you do not get from being taught, or from books, etc. but get from personal experience, for example when working in a particular organization. See https://dictionary.cambridge.org/dictionary/english/tacit-knowledge .

[18] Rules for Domain names, see https://eurid.eu/en/register-a-eu-domain/rules-for-eu-domains/

[19] Habibzadeh P. Decay of references to Web sites in articles published in general medical journals: mainstream vs small journals. Appl Clin Inform. 2013 Oct 2;4(4):455-64. doi: 10.4338/ACI-2013-07-RA-0055. PMID: 24454575; PMCID: PMC3885908.

[20] Loan, F.A. and Shah, U.Y. (2020), "The decay and persistence of web references", Digital Library Perspectives, Vol. 36 No. 2, pp. 157-166. https://doi.org/10.1108/DLP-02-2020-0013

[21] The handle system is defined in https://www.rfc-editor.org/rfc/rfc3650.txt, https://www.rfc-editor.org/rfc/rfc3651.txt and https://www.rfc-editor.org/rfc/rfc3652.txt

[22] The Handle Registry at http://hdl.net/

Object Identifier[23] (DOI), or others such as ARK[24], PURL[25] depend upon servers, all require resources, and maintenance to update the new location of the resources if/when they are re-located.

1.4.5 Slow changes

We tend to believe that change happens relatively slowly, so that we will have time to adapt to the changes. However, clearly disasters, such as floods, earthquakes and violent storms do happen. Companies, even very large ones, suddenly stop operating, through bankruptcy or legal issues. Criminals or terrorists or nations can strike unexpectedly.

A sensible precaution is to prepare to mitigate such events, such as not building on flood plains or seismic faults or taking out insurance or installing burglar alarms.

However, there is a famous quote from Donald Rumsfeld:[26] *"…there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know*."

We must beware of the "unknown unknowns" that cause rapid changes that we cannot mitigate but to which we must react. To be able to do this one must be sure we can understand systems, and bits etc, in depth.

1.5 Common errors in digital preservation

At this point it is worth thinking about some common errors when digital preservation is discussed. Some are obvious but others are more subtle and are errors which are often overlooked. This list is not exhaustive but should spark some thoughts in the reader.

People can get so confused by many things including:

- the minutiae of their jobs mean that they forget the fundamentals;
- others tell them that digital preservation is difficult and so one must accept the provided tools, recipes, and cookbooks without understanding their limitations;
- not thinking clearly and logically.

People get fooled by:

- the software that does things "auto-magically" for you when you click on a file;
- examples that look backwards rather than forward;
- people who claim to know about digital preservation, but actually don't;
- people, especially vendors, who speak only about **formats** but never speak about semantics or information.

People tend to forget:

- what their current software does for them now;
- that software can disappear;
- that things tend to get more complex and interlinked;
- that they will need to know how to preserve things that are not just for display or rendering;
- that digital preservation is about preserving the information that is encoded in bits;
- that bits can mean **anything**.

What people don't usually think about:

[23] DOI standard https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en

[24] See https://datatracker.ietf.org/doc/html/draft-kunze-ark-01

[25] See https://purl.archive.org/

[26] See https://www.youtube.com/watch?v=REWeBzGuzCc

- the details of the bits and the potential complications inherent in digital objects.

What people tend to panic about:

- complexity and the unknown.

1.6 Limitations of this chapter

This chapter clearly can only cover a limited number of aspects related to digital preservation but should provide the basic ideas on which to build.

2 What OAIS says and why it says it

If language is not correct, then what is said is not what is meant; if what is said is not what is meant, then what must be done remains undone; if this remains undone, morals and art will deteriorate; if justice goes astray, the people will stand about in helpless confusion. Hence there must be no arbitrariness in what is said. This matters above everything[27].

*(**Confucius** 551 BC - 479 BC)*

This chapter aims to provide the basic ideas and concepts needed to build the rest of this book on. We do this by jumping in feet first, based on the terminology from the OAIS Reference Model, based on the ideas introduced in earlier chapters. We need to do this in order to be able to talk clearly about digital preservation, because we want to say what we mean.

2.1 The aims of OAIS

As we (the CCSDS DAI Working Group[28], and before that Panel 2, which I lead) wrote OAIS[29] we had several aims and guiding principles. The group consists of people with various backgrounds including libraries, national archives, and scientific data repositories. We realized that there were deficiencies in the previous work on digital preservation, namely:

- a focus on the document world, with concepts that were not applicable to scientific data;
- existing vocabularies were unclear:
    - different communities used different words for very similar concepts;
    - different communities used the same word for very different concepts;
- it was not clear how to assess any of the concepts, in particular how to know whether digital preservation had been, or was likely to be, successful.

We decided to avoid:

- things to do with funding;
- things to do with speed of response to requests;
- things that are tied to specific organisations;
- legal issues arising in any specific judicial system;
- details of designs for a repository;
- adding every possible idea into the document;
- being tied to "Open Data";

[27] See http://www.quotationspage.com/quote/14176.html

[28] CCSDS DAI working group page https://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI

[29] The definitions used here are from the updated version of OAIS. The details of the changes from the OAIS 2012 version are provided in section 4

- being tied to specific hardware or software

We realized that we must have a standard which:

- had concepts that are applicable to any digital object;
- had concepts that are testable;
- defined what was required for conformance, although we realized that this standard itself could not cover everything required of an archive;
- defined various roles of actors, whether people, organisations, or systems, recognizing that a specific actor could take on various roles at different times;
- defined a roadmap of follow-on standards;
- had concepts that would allow the creation of a certification system;
- did not assume that anyone could see into the future, but instead defined processes which were needed to carry out digital preservation as things change;

These considerations led us to produce a standard which:

- was not a design for a repository, although some readers do read the standard as a guide or checklist for a design;
- was not specific about HOW to do things;
- did start from very fundamental concepts, providing a logical sequence of ideas - beginning with a number of definitions;
- was applicable to any digital holdings;
- was testable;

It had some parts are MANDATORY for OAIS conformance, and these are:

  - carefully worded
  - practical
  - minimal

Other parts provide terminology and what we thought would be useful guidance.

A key decision was to say nothing about a topic unless we could say something useful.

2.2 Introduction to OAIS concepts and terminology

Another way of looking at this is to realise that different people have slightly different definitions in mind, depending upon their backgrounds, for many common terms. If we are not careful we will talk at cross-purposes because of these differences. In order to avoid this, we need clear definitions.

The next few sections discuss some of the basic OAIS definitions and concepts.

2.3 Preserve what, for how long and for whom?

The “O” in OAIS stands for “Open” but refers to the open way the standard was developed rather than anything to do with open access. Indeed, the OAIS Reference Model can apply to any type of archive whether open access, closed, restricted, “dark” or proprietary.

OAIS takes a very general definition of its prime concern which, as the “I” in OAIS suggests, is information:

Information : ***Any type of knowledge that can be exchanged. In an exchange, it is represented by data.*** An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius.

Note that Knowledge is not defined in OAIS; the word is used in the normal dictionary meaning:

- Cambridge dictionary[30]
  - “understanding of or information about a subject that you get by experience or study, either known by one person or by people generally
- Collins[31]:
  - “Knowledge is information and understanding about a subject which a person has, or which all people have.”
- Miriam-Webster:
  - A
    1) the fact or condition of knowing something with familiarity gained through experience or association
    2) acquaintance with or understanding of a science, art, or technique
  - B
    1) the fact or condition of being aware of something
    2) the range of one's information or understanding

2.4 OAIS Information Model component definitions

The accompanying definition of data is equally broad:

Data : ***A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.*** Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.

There are related definitions:

Data Object: Either a Physical Object or a Digital Object.

Physical Object: An object (such as a moon rock, bio-specimen, microscope slide) with physically observable properties that represent information that is considered suitable for being adequately documented for preservation, distribution, and independent usage.

In the case of things digital, OAIS defines:

Digital Object : ***An object composed of a set of bit sequences***.

Note that this does not mean we are restricted to a single file. The definition includes multiple, perhaps distributed, files, or indeed a set of network messages.

The restriction to “bits” i.e., consisting of “1” and “0”, means that if we move to trinary (i.e., “0”, “1” and “2”) or quantum bits instead of binary then we would have to change this definition, but it would not affect the concept – however it would change the tools we could use.

One might wonder why data includes physical objects such as a "moon rock specimen". The answer should become clear later but in essence the answer is that to provide a logically complete solution to digital preservation one needs, eventually, to jump outside the digital, if only, for example, to read the label on the disk, or a copy of the ASCII encoding carved in stone .

As to the question of length of time we need to be concerned about, OAIS provides the following pair of definitions (the text in bold italics below is taken from OAIS):

Long Term : ***A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a***

[30] See https://dictionary.cambridge.org/dictionary/english/knowledge

[31] See https://www.collinsdictionary.com/dictionary/english/knowledge

changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.

Long Term Preservation : ***The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.***

In other words, we are not only talking about decades into the future but, as is a common experience, we need to be concerned with the rapid change of hardware and software, the cycle time of which may be just a few years. Of course, even if an archive is not itself looking after the digital objects over the long term, even by that definition, the intention may be for another archive to take over later. In this case the first archive needs to capture all the *"metadata"* needed so that it can hand these on also.

Note that OAIS does use the word "metadata" because it is too broad a term, as discussed previously, but we use the term, carefully, in this book for convenience.

Three key concepts are embedded in the above definition namely:

Authenticity: ***the degree to which a person (or system) may regard an object as what it is purported to be. The degree of Authenticity is judged on the basis of evidence.***

There will be much more to say about authenticity in section 3.3.6 and chapter 4, where the whole chapter is devoted to it.

Independently Understandable : ***A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community, as exemplified by the associated Preservation Objectives, without having to resort to special resources not widely available, including named individuals***.

This introduces another concept "Preservation Objectives" which we come to next.

By being able to "understand" a piece of information is meant that one can do something useful with it; it is not intended to mean that one understands all of its ramifications.

For example, in a criminal investigation of a murder one may have a database with digitally encoded times of telephone calls; here we would be satisfied if we could say "the telephone call was made at 12:05 pm on 1st January 2009, UK time", but to then understand that this implied that the person who made the call was the murderer is beyond what OAIS means by being able to "understand" the data.

In OAISv3 we strengthened the concept of understandability by making it testable. We did this by defining:

Preservation Objectives: **A specific achievable aim which can be carried out using the Information Object.**

Preservation Objectives are intended to allow the repository to make it possible to test whether the information actually is Independently Understandable by members of the Designated Community now and into the future, in particular having adequate Representation Information. In order to facilitate this each Preservation Objective should be:

- Specific – The objective should be well defined and clear to anyone with the assumed Knowledge Base
- Actionable – The objective should be achievable currently and into the future.
- Measurable - It should be possible to know whether or not the objective has been attained at a given point in time.

Examples of Preservation Objectives include:

- The ability to render documents, images, videos, or sounds in a way which is sufficiently similar to the original. This could be checked by verifying that, for example, the document is readable, or the image is viewable. An analysis of the

colours could also be compared. A spectral analysis could be performed on the sounds and compared with that of the original.
- The ability to process a dataset and generate the data products expected. This could be checked by comparing with something generated earlier, for example on Ingest.
- The ability to understand a dataset and use it in analysis tools to generate results, for example the density of electrons in the upper atmosphere or the structure of a molecule, given certain measurements. These could be compared with results generated earlier.
- The ability to re-perform an artistic performance. This could be compared with a recording of a previous performance.

Checks on the success of the preservation activity should include confirmation that these aims have been fulfilled. Clearly information needed to make comparisons, as in the examples above, would be expected to be created earlier, perhaps on Ingest, and would need to accompany the information being preserved.

An OAIS archive must define Preservation Objectives. The definition of the Preservation Objectives may be subject to agreement with funders and other stakeholders. They will likely provide input into those definitions and may have agreements with the OAIS regarding those definitions. The definition of the Designated Community and its Knowledge Base will probably change over time and therefore the definition of the Preservation Objectives may change over time. In this case, funders and other stakeholders would again be consulted.

Now we approach one element of what that the "preservation" part of "digital preservation" means. To require that things are able to be "interpreted, understood and used" is to make some very powerful demands. It not only includes playing a digital recording so it can be heard, or rendering an image or a document so that it can be seen; it also includes being able to understand what the columns in the spreadsheet we mention earlier means, or what the numbers in a piece of scientific data mean; this is needed in order to actually understand and in particular **use** the data, for example using it in some analysis programme, combining it with other data in order to derive new scientific insights. The "Independently" part is to exclude the easy but unreliable option of being able to simply ask the person who created the digital object; unreliable not because the creator may be a liar but rather because the creator may be, and in the very long term certainly will be, deceased!

Finally, we have the other key concept of “Designated Community”.

: ***An identified group of potential Consumers who should be able to understand a particular set of information in ways exemplified by the Preservation Objectives. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the archive and this definition may change over time.***

Why is this a key concept? To answer that question, we need to ask another fundamental question, namely "How can we tell whether a digital object has been successfully preserved?" – a question which can be asked repeatedly as time passes. Clearly we can do the simple things like checking whether the bit sequences are unchanged over time, using one or more standard techniques such as digital digests. However, just having the bits is not enough. The demand for the ability for the object to be "interpreted, understood and used" is broader than that - and of course it can be tested.

The word "should" means that the repository is saying that it commits to make sure that the members of the Designated Community will be able to understand.

But surely there is another qualification, for is it sensible to demand that *anyone* can "interpret, understand and use" the digital object - say a four-year-old child? Clearly we need to be more specific. But how can such a group be specified, and indeed who should choose? This seems a daunting task - who could possibly be in a position to do that?

The answer that OAIS provides is a subtle one. The people who should be able to "interpret, understand and use" the digital object, and whom we can use to test the success or otherwise of the "preservation", are defined by the people who are doing the preservation.

The advantage of this definition is that it leads to something that can be tested. So, if an archive claims "we are preserving this digital object for astronomers" we can then call in an astronomer to test that claim, using the Preservation Objectives.

The disadvantage is that the preserver could choose a definition which makes life easy for him/her – what is to stop that? The answer is that there is nothing to prevent that BUT who would deposit their information in an archive for which the definition does not suit the depositor? The main point is that the archive must make its definition clear so that people can judge.

As long as the archive's definition is made clear then the person depositing the digital objects can decide whether this is acceptable. The success or failure of the archive in terms of digital objects being deposited will be determined by the market. Thus, in order to succeed the archive will have to define its Designated Community (ies) appropriately.

Members of the Designated Community should be identifiable so that the repository can test whether it has enough Representation Information by asking some of those people.

There has been some misunderstanding[32] about the use of the term "potential Consumers" in the definition of Designated Community, where it has been taken to mean that the Consumers are in some sense non-existent. Rather it means that any specific individual in the Designated Community may not actually be a user, but they could be. For example, if the Designated Community for a particular dataset is "astronomers", if we select a specific astronomer he/she may not actually be a Consumer i.e., may not actually get the information from the Archive, but he/she is a "potential Consumer" i.e., he/she could get the information from the Archive.

OAIS says:

> *For example, an Archive may decide that certain Content Information should be understandable to the general public and, therefore, this becomes the Designated Community.*
>
> *For some scientific information, the Designated Community of Consumers might be described as those with a first-year graduate level education in a related scientific discipline. This is a more difficult case as it is less clear what degree of specialized scientific terminology might actually be acceptable. The Producers of such specialized information are often familiar with a narrowly recognized set of terminology, so it is especially critical to clearly define the Designated Community for that information and to make the effort to ensure that this community can understand the information.*
>
> *The possible changes to the definition of the Designated Community also need consideration. Information originally intended for a narrowly defined community may need to be made more widely understandable at some future date. For example, information originally intended to be understandable to a particular scientific community may need to be made understandable to the general public. This is likely to mean adding explanations in support of the Representation Information and the Preservation Description Information, and it can become increasingly difficult to obtain*

[32] Bettivia, R.S. (2016), The power of imaginary users: Designated communities in the OAIS reference model. Proc. Assoc. Info. Sci. Tech., 53: 1-9. https://doi.org/10.1002/pra2.2016.14505301038

> *this information over time. Selecting a broader definition of the Designated Community (e.g., general public) when the information is first proposed for Long Term Preservation can reduce this concern and also improve the likelihood that the information will be understandable to all in the original community.*

ISO 16363 provides additional examples of Designated Community:

- General English-reading public educated to high school and above, with access to a Web Browser (HTML 4.0 capable).
- For Geographic Information System (GIS) data: GIS researchers—undergraduates and above—having an understanding of the concepts of Geographic data and having access to current (2005, USA) GIS tools/computer software, e.g., ArcInfo (2005).
- Astronomer (undergraduate and above) with access to Flexible Image Transport System (FITS) software such as FITSIO, familiar with astronomical spectrographic instruments.
- Student of Middle English (a form of the English language spoken after the Norman conquest (1066) until the late 15th century) with an understanding of Text Encoding Initiative (TEI) encoding and access to an XML rendering environment.
    - Variant 1: Cannot understand TEI;
    - Variant 2: Cannot understand TEI and no access to XML rendering environment;
    - Variant 3: No understanding of Middle English but does understand TEI and XML.
- The repository has defined the external parties, and its assets, owners, and uses. Two groups: the publishers of scholarly journals and their readers, each of whom have different rights to access material and different services offered to them.

Some repositories may have a policy not to allow consumers to get access to its contents for a certain period of time call themselves, sometimes called a ‘dark archive’, but they would nevertheless need a Designated Community.

Different archives, holding the same digital object may define their Designated Communities as being different. This will have implications for the amount and type of *“metadata”* which is needed by each archive.

There are publications[33] which use the term “Designated User Community” rather than “Designated Community”. While such a term is understandable, nevertheless the term is not used in OAIS and so is best avoided because it could be misleading.

2.4.1 What *“metadata”*, how much *“metadata”*?

One fundamental question to ask is ‘What *“metadata”* do we need?’ The problem with *“metadata”* is that it is so broad that people tend to have their own limited view. OAIS provides a more detailed breakdown. The first three broad categories are to do with (1) understandability, (2) origins, context, and restrictions and (3) the way in which the data and *“metadata”* are grouped together.

The reason for this separation is that, given some digitally encoded information, one can reasonably ask whether it is usable, which is dealt with by (1). This is a separate question to

[33] See for example M. A. Parsons, R. Duerr, Designating user communities for scientific data: challenges and solutions, Data Science Journal, 2005, Volume 4, Pages 31-38, Released on J-STAGE January 05, 2006, Online ISSN 1683-1470, https://doi.org/10.2481/dsj.4.31, https://www.jstage.jst.go.jp/article/dsj/4/0/4_0_31/_article/-char/en

the one about where this digital object came from, dealt with by (2). Since there are many ways of associating these things it seems reasonable to want to consider (3) separately.

It could be argued that to understand a piece of data one needs to know its context. However, the discussion about “Independently Understandable” in the previous section points out that OAIS does not require understanding of all the ramifications so this separation of context from understandability is reasonable, although it does **not** mean that all context is excluded from understandability since a piece of *“metadata”* may have several roles.

The packaging is something which is separate from the content. The next few sub-sections briefly introduce these different categories; they will each be discussed in much greater detail in separate chapters.

2.4.1.1 Understandability (Representation Information)

One type of *“metadata”* we can immediately identify is that which we need to "interpret, understand and use" the digitally encoded information. OAIS defines this as:

Representation Information : ***The information that maps a Data Object into more meaningful concepts so that the Data Object may be understood in ways exemplified by Preservation Objectives.*** An example of Representation Information for a bit sequence which makes up a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning of keywords in the file which are not part of the standard.

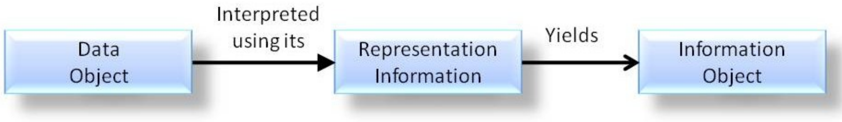


Figure 5 Representation Information

Figure 5 indicates that the Representation Information is used to interpret the Data Object in order to produce the Information Object – something which one can then understand and use.

The OAIS definition of **Information Object** is: ***A Data Object together with its Representation Information.*** This is a very broad definition.

The definition of Information Object may seem a little circular. However, its purpose is not to define something specific, for example in a computer programme. Instead, it really only provides a simple term for something which we can apply to many different things in people’s heads. The key idea is that it is something that allows us to talk about what knowledge is being exchanged.

When we are referring to something specifically targeted for preservation the term **Content Information** is used. This is ***a set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information.*** Some might think that the “original target of preservation” must be something like a piece of scientific data, but a better understanding is that it means the object we are focusing on right now. We can focus on the other objects in this diagram, and later diagrams, such as the Representation Information, later, when it will be regarded as ”the target of preservation” and referred to as Content Information.

In a little bit more detail, recognising that the Data Object could be either digital or physical, one can draw Figure 6, which is a simple UML[34] diagram.

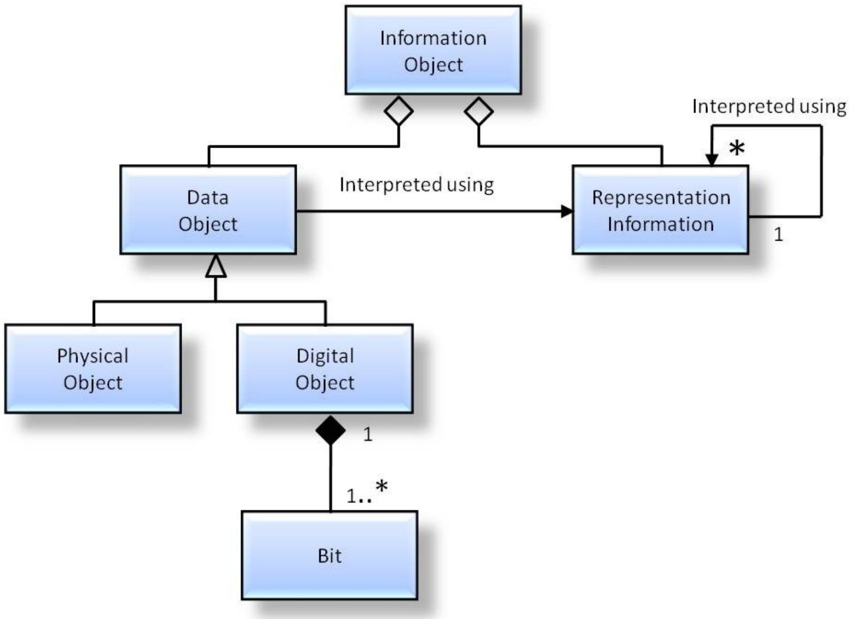


Figure 6 OAIS Information Model

This diagram is a way of showing that

- an **Information Object** is made up of a **Data Object** and **Representation Information**
- a Data Object can be either a **Physical Object** or a **Digital Object**. An example of the former is a piece of paper or a rock sample.
- a Digital Object is made up of one or more **Bits**

Note that this does not mean we are restricted to a single file. The definition includes multiple, perhaps distributed, files, or indeed a set of network messages.

- a Data Object is interpreted using Representation Information

It is important to realise that Representation Information can be anything from a scribbled handwritten note, needing a human to read it, to a complex machine-readable formal description.

- Representation Information is itself interpreted using further Representation Information

It should be borne in mind that any piece of Representation Information may be found to be erroneous in future or may be improved by a better understanding of the data object or may be completely replaced.

2.4.2 Recursion – a pervasive concept

Those with a mathematical background will recognise some of this as a type of recursion. It comes up time and again in preservation. By this we mean that ideas which appear at one level of granularity re-appear when we take a finer grained view, within the detailed breakdown of those or other ideas. As is well known in mathematics, it is important to understand where the recursion ends otherwise it becomes impossible to produce practical

[34] See https://www.uml.org/ and Annex C of the OAIS standard

results. For example, the factorial function is defined as n! = n*((n-1)!) i.e., 6! = 6*(5!) = 6*5*(4!) = ... This stops when we get to 0! because we define 0! as equal to 1.

It is worth making some remarks about this concept here.

Representation Information (RepInfo for short) – remember it is Representation Information rather than Representation Data - is encoded as data (which could be called representation data but in fact OAIS does not use that terminology) which itself needs its own Representation Information. The recursion stops at the Knowledge Base of the Designated Community.

Any piece of “metadata”, such as Provenance (to be discussed in detail later), will itself be encoded as a Data Object, which needs Representation Information. Representation Information as a digital object will also need its own Provenance, as illustrated in Figure 7

The recursion in this case might end with Provenance being a simple text file (or piece of paper) in plain English (assuming the Designated Community can read English) so the Representation Information is quite simple and hence the Representation Information Network terminates.

In OAIS many of the concepts that are used are Information Objects.

As noted previously, each of the items of PDI and anything else which an archive seeks to preserve, will be an Information Object, that is it is encoded in a Data Object, which has Representation Information associated with it.

But that Representation Information will also have some Provenance associated with it as illustrated in Figure 7.

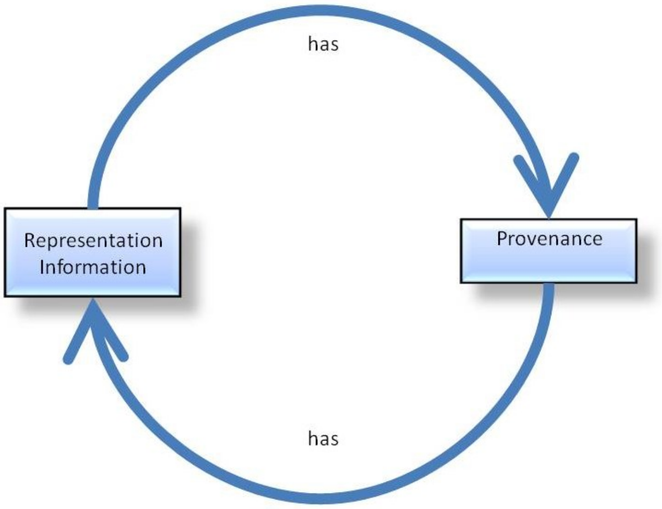


Figure 7 Recursion - Representation Information and Provenance

Thinking about every element which the archive seeks to preserve, besides the Representation Information for the Data Object, it must surely know from where it got that Data Object, and what has happened to it since receipt, it must be confident that the Data Object has not been changed; it must know how to find it; it must know what the context of that Data Object is and finally it must be clear about what the access restrictions are on that Data Object, if any.

Standing back and comparing these considerations with the objects which are the “main preservation focus” of the archive, then all these “metadata” elements must themselves be associated with all the elements needed for an Archival Information Package (see section 2.4.4.2)..

The archive will need to preserve many pieces of information besides the AIPs. The following lists show items mentioned in this document, but the archive may preserve other items.

2.4.2.1.1 Subtypes of Representation Information

Figure 8 denotes that Representation Information may usefully be sub-categorised into several different types, namely Structure, Semantic and (the imaginatively named) Other Representation Information. This breakdown is useful because Structure Representation Information is often referred to as “format”; Semantic Representation Information covers things such as ontologies and data dictionaries; Other Representation Information is a catch-all for anything and everything else.

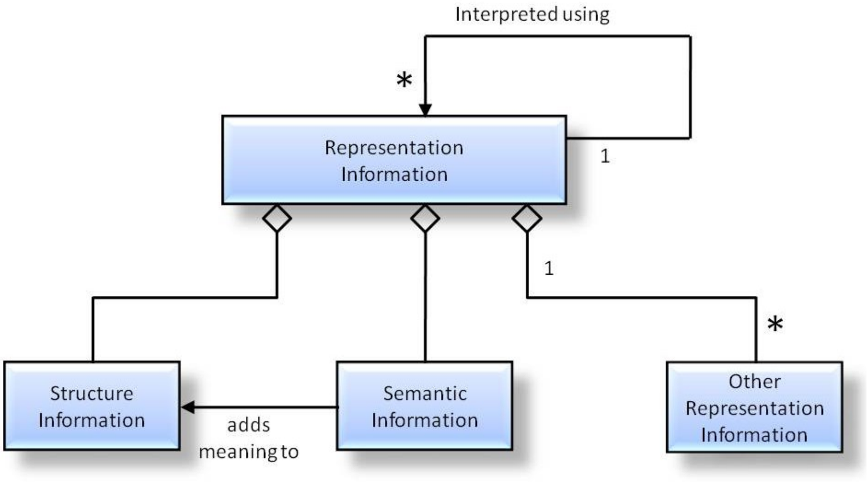


Figure 8 Representation Information object

One useful way to understand why this breakdown may be useful is to consider a number of different variations.

For example, two copies of a simple message (i.e., a piece of information) may be contained in two text files (i.e., in the same format), but in one case the message is written in English and in the other case it is in French (needing different dictionaries).

Similarly, one can have the English text both in a PDF and a Word file – two different formats but needing the same dictionary.

In general, breaking things down into smaller pieces means that one is not forced to treat objects as a sticky mess. Instead, one can deal with each (smaller) part separately and usually more easily.

When this is coupled with the fact that Representation Information is an Information Object that may have its own Data Object and other Representation Information associated with understanding that Data Object, as shown in a compact form by the *interpreted using* association, the resulting set of objects can be referred to as a Representation Network.

In the extreme, the recursion of the Representation Information will ultimately stop at a physical object such as a printed document (ISO standard, informal standard, notes, publications etc). This allows us to make a connection to the non-digital world. However, use of things like paper documentation would tend to prevent "automated use" and "interoperability", and also complete resolution of the complete Representation Network, discussed further below, to this level would be an almost impossible task. Therefore, we would prefer to stop earlier, and this will be discussed next.

As the final part of this tour through the OAIS concepts we turn to something a little different in order to answer the question '***How much "metadata"?***'

A piece of Representation Information is just another piece of Information – hence the name Representation Information rather than Representation Data. In order for there to be enough Representation Information it has to be understandable and usable by the Designated Community - in order to be used to understand the original data object. However, what if this is not the case?

The Representation Information may be encoded as a physical object such as a paper document, or it may be a digital object. In the latter case we can simply provide Representation Information for that digital object. If the Designated Community still cannot understand and use the original data, we can repeat the process. Clearly this provides us with a way to answer the "How much" question: we provide a network of Representation Information until we have enough for the Designated Community to understand the Data Object. OAIS defines:

Representation Network: ***The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term***.

To complete the picture, we can then see a way to define the Designated Community, namely we define them by what they know, by what OAIS terms their Knowledge Base:

Knowledge Base : ***A set of information, incorporated by a person or system that allows that person or system to understand received information.***

2.4.3 Preservation Description Information

2.4.3.1 Besides Understandability – are we sure it is what we think it is?

Let us think what else one needs for preservation of some digitally encoded information, beyond understanding it.

Clearly one needs to be sure that

- If one has some kind of identifier, then we get the right object;
- one needs to know that the object has not been altered in some unexpected way;
- one needs to know that the object is what we think it is;
- one needs to know how the object is related to other things;
- one needs to know that the object can be dealt with legally

Preservation Description Information (PDI) collects together this information under several headings as follows. This information, which along with Representation Information, is necessary for adequate preservation of the Content Data Object and which can be categorized as Provenance Information, Context Information, Reference Information, Fixity Information, and Access Rights Information. It is a type of Information Object.

Defining (as well as its components: Provenance Information, Context Information, Reference Information, Fixity Information, and Access Rights Information) as relevant to the Content Data Object does not mean that those concerns are any less important for other data objects or at other levels, for example, it is important to apply reference, fixity, provenance, context, and access rights to Representation Information, or to any other information the Archive is preserving. Definition of these terms as relevant to the Content Data Object is simply to ease discussion of these concepts at the Content Data Object level.

For each of the components discussed next is a type of Information Object.

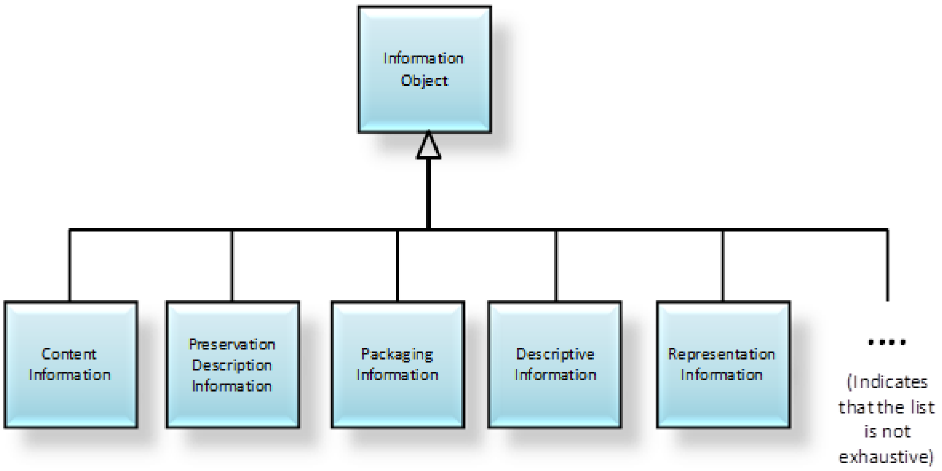

Information Object
Content Information
Preservation Description Information
Packaging Information
Descriptive Information
Representation Information
....
(Indicates that the list is not exhaustive)

Figure 9 Information Object Taxonomy

However, one may immediately think of, for example, Fixity Information as a hash , but that hash is the Data Object (some bits) which requires its own Representation Information in order to be an Information Object. The Representation Information could tell us that the bits encode characters using ASCII-7, and the hash algorithm used was SHA256. This was discussed in section 2.4.2.

2.4.3.2 Reference Information

Reference Information: The information that is used as an identifier for the Content Data Object. It also includes identifiers that allow outside systems to refer unambiguously to a particular Content Data Object.

It identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Data Object. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Data Object.

Examples of these systems include taxonomic systems, reference systems and registration systems. In the OAIS Reference Model most if not all of this information is replicated in Package Descriptions, which enable Consumers to access Information of interest.

The Data Object of the Reference Information may again be some bits and the Representation Information could be that this is a UNICODE[35] encoding using a particular code point. Using this one might then see it is "urn::xyx::abc::123" – in order to understand this one would need Semantic Representation Information which explains how to use this set of characters to find the object.

2.4.3.3 Fixity Information

Fixity Information: The information which documents the mechanisms that ensure that the Content Data Object has not been altered in an undocumented manner.

Fixity Information provides the data integrity checks or validation/verification keys used to ensure that the particular Content Data Object has not been altered in an undocumented manner. Fixity Information could include special encoding and error detection schemes that are specific to instances of Content Data Objects.

Fixity Information documents the mechanisms that ensure that the Content Data Object has not been subject to undocumented alteration. Examples include Cyclic Redundancy Check

[35] See for example https://www.smashingmagazine.com/2012/06/all-about-unicode-utf8-character-sets/

(CRC) codes, checksums, or message digests. The mechanisms are not excluded from being used for other objects, but the mechanisms used for Content Data Objects are of particular interest.

It is important to understand that while, for example, hashes or digests of digital object may sometimes be referred to as Fixity, OAIS means that one must know how the system works. For example, when we calculate a hash and compare it to the "original" hash, how can we be sure that the "original" hash has not been changed?

2.4.3.4 Provenance Information

Provenance Information: The information that documents the history of the Content Data Object. This information tells the origin or source of the Content Data Object, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity.

Provenance Information documents the history of the Content Data Object. This tells the origin or source of the Content Data Object, checks on its Information Properties to be preserved (Transformational Information Properties), any changes that may have taken place since it was originated, and who has had custody of it since it was originated, providing an audit trail for the Content Data Object. This gives future users some assurance as to the likely reliability of the Content Data Object as it contributes to evidence supporting Authenticity. Provenance Information can be viewed as a special type of Context Information.

The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity.

The Data Object of the Provenance Information may be a set of sequences of bits which is the Structure Representation Information tells us is a PREMIS file (see section 1) encoded in EBCDIC[36], and using a specialised vocabulary specified in the Semantic Representation Information.

2.4.3.5 Context Information

Context Information: The information that documents the relationships of the Content Data Object to its environment. This includes why the Content Data Object was created and how it relates to other Content Data Objects.

The Data Object of the Context Information could be some bits for which the Structure Representation Information is that it is a set of characters encoded in ASCII-7. The Semantic Representation Information could be that this is "UK English" text using a specific special terminology.

2.4.3.6 Access Rights Information

Access Rights Information: The information that identifies the access restrictions pertaining to the Content Data Object, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures.

[36] Extended Binary Coded Decimal Interchange Code (EBCDIC) - see https://www.ibm.com/docs/en/db2-for-zos/12?topic=schemes-ebcdic

It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures.

2.4.3.6.1 Subtleties of Access Rights

When one hears about Digital Rights, one will probably think about restrictions and payment of fees that one must respect if one wants to download and enjoy one's favourite song or read some parts of the intriguing e-book about digital preservation found on Internet. That's true, but Digital Rights exist and have a legal validity even if one is not forced to respect the conditions. So, which are the issues that Digital Rights pose on the long-term preservation?

If one is preserving "in-house" all the pictures taken since the purchase of a digital camera, then there is no problem. But if one needs to curate of some artistic, cultural, or scientific material that was produced by someone else, then the law normally imposes limitations on the use, distribution, and any kind of exploitation of that material.

You might think "Fine, I know already what I'm allowed to do! Why should I further care about rights?" The reason is that things will change: new laws will come into force and the Copyright will, at a given time in the future, expire or the heirs of the original rights holder could give up the exploitation rights and put the work being preserved into the Public Domain. All these things have an impact on what anybody is allowed to do.

Is there anything else to think about, except Copyright? Yes, there is Protection of Minors, Right to Privacy, Trademarks, Patents, etc., and they all share the same aim: they protect people from potential damages due to incorrect use of the material being held! One should be aware of that.

The main questions that must be asked are:

- do the activities related to digital preservation violate any of the above rights?
- are there some limits in copying, transforming, and distributing the digital holdings?
- is the object of preservation some personal material or is it intended for a wider public?

Future consumers will have to respect the same limitations, and they should also be informed about the special permissions that the Laws grant them or that the rights holder was willing to grant. In other words, access conditions depend both on legislation and on conditions defined within licenses and both must be preserved over time and be kept updated.

2.4.3.6.1.1 Limitations and rights to perform digital preservation

Preserving a digital work in the long-term requires that a number of actions are undertaken, including copying, reproducing, making available and transforming its binary representation.

These actions might infringe existing Copyright: for instance, if one wanted to transform a digital object from an obsolete format to a most recent one, and thereby risk altering the original creation in a way that the rights holder might not agree with.

To ascertain that no such exclusive rights are violated, a preservation institution has the following main options (which are all, within the conditions defined, in line with the OAIS mandatory responsibilities):

- to become the owner of the digital material and to obtain the exclusive rights from the creators (excluding the non-transferrable moral rights);
- to preserve only material that is in Public Domain (e.g., where Copyright is expired, or the author has released the work into Public Domain);

- to carry out preservation in accordance with the conditions defined by the Law (e.g., in some countries there are Copyright Exceptions which grant some kinds of institutions the permissions to perform digital preservation)
- to obtain from the rights holders, by means of a license, the permissions to carry out the necessary preservation activities.

Many countries have defined exceptions in their Copyright Laws to facilitate libraries, archives, and other institutions to carry out digital preservation. However, until a legal reform is carried out, it is good practice to get the required authorization from the rights holders through rights transfer contracts or licenses, and not to rely solely on the existing jurisdiction to ensure a comprehensive preservation of copyrighted materials.

2.4.3.6.2 Preserving limitations and rights over time

At some time in the short or long term, somebody will desire or need to access one of the preserved archive holdings. Protection of Minors and Privacy Laws regulate the use of particular types of data. However, the most complex limitations come from Intellectual Property Rights (IPRs): Copyright, Related Rights, and Industrial Property Rights, such as Trademarks, Industrial Design and Patents.

Dealing with IPR-protected material poses risks, because it could conflict with the normal exploitation of the work or prejudice the legitimate interests of the rights holders. Therefore, the preservation institution should reduce the risk taken by future consumers and try to arrange things so that those consumers are able lawfully to exploit the materials.

We will see that it is not enough just to identify and store the details on who holds some Copyright and the licenses that are attached to the content; it is also necessary to preserve other kinds of information, to monitor the changes in the legislation and to be continuously updated about the ownership of rights. If the consumers were authorized to exploit a piece of content in the way they intend, they should have the ability to show the appropriate authorization.

Since the revision of the OAIS Reference Model a specific section of the Preservation Description Information (PDI) has been defined to address authorization in the long-term, namely Access Rights Information. This information is specified in part by the rights holders within the Submission Agreement. For example, it could contain the license to carry out preservation activities, licenses offered to interested consumers and the rights holders' requirements about rights enforcement measures. But this PDI section could even include the special authorizations that are granted by the law. In short, OAIS Access Rights include everything related to the terms and conditions for preservation, distribution, and usage of the Content Information.

There are two kinds of access rights to be considered. On the one hand there are the exclusive ownership rights that are typically held by the owners of the works, and on the other hand there are the non-exclusive permissions that are granted to other persons. In order to be able to correctly preserve all the existing rights - exclusive ownership rights and non-exclusive permissions - the following information is required:

- Ownership of rights
- Licences
- Rights-relevant Provenance information
- Post-publication events
- Laws

Each of these is discussed in turn below.

2.4.3.6.2.1 Ownership of rights

Ownership rights can be derived from the application of the Law to provenance and to post-publication events. Thus, one could just preserve the latter and “calculate” the existing rights only when the legitimacy of some intended action must be controlled.

In practice however it is useful to have the ownership rights already processed and stored in explicit form, for instance for statistical purposes and for searching and browsing the preserved material. This requires that adequate mechanisms are put in place for notification about changes in the Law and on some other relevant events in the history of a work, because these could imply some change in the status of rights.

2.4.3.6.2.2 Licenses

When rights holders are willing to grant some specific permission to other people to exploit their creation, they can do this through a licence. Licences contain the terms and conditions under which the use of the creation is permitted.

Preserving licences over time gives the future consumer a better chance to exploit an intellectual work.

2.4.3.6.2.3 Rights-relevant Provenance Information

This information includes the main source of information from which the existing exclusive rights can be derived by applying the Law. In the simplest case it corresponds to the creation history, saying who the creators are, when and in which country the creation was made public for the first time, and the particular contribution of each creator.

However, the continuously changing legislation poses a challenging issue, namely that it is impossible to predict which information might be relevant.

Consider for example that France has, at a certain point, extended the Copyright duration with provision of five and nine years respectively for works created in the years of the First and the Second World War, and it has added further thirty years if the author "died for France"[37]. This means that the publication year is not sufficient to derive the rights, as it is necessary also to trace if an author died during active service!

This kind of information is absolutely crucial to correctly identify all the existing ownership rights, their duration and the jurisdiction under which they are valid.

2.4.3.6.2.4 Post-publication events

This information concerns events that have an impact on ownership rights and on permissions, but which cannot be considered as part of the creation history. It includes:

- Death of a creator: the date of death influences the duration of the ownership rights; the identities of the heirs are crucial if particular authorizations need to be negotiated
- Release in Public Domain: the rights holders might decide to give up all rights even before the legal expiration date
- Transfer of Rights: the rights holders might transfer some or all of their exclusive rights to someone else.

If this kind of information is preserved and kept updated, it should be possible to exploit the IPR-protected material in the near and the far future.

2.4.3.6.2.5 Laws

Tracking laws is crucial for the correct preservation of rights: changes must be immediately recognized because they might strengthen or reduce the legal restrictions for some materials.

[37] Right might still be wrong, 2013, see https://www.kl.nl/opinie/right-might-still-be-wrong/

Laws need not to be preserved themselves, but an archive should be able to recognize and to handle the changes. This is true not only for Intellectual Property Rights, but also for Right to Privacy and Protection of Minors.

2.4.3.6.3 Rights enforcement technologies

Technological solutions like encryption, digital signatures, watermarking, fingerprinting and machine-understandable licenses could be applied to enforced access rights. Thus, the rights holders and content providers could ask the preservation institution to make the deposited material available only under some restrictions and to enforce them with proper security measures.

Each OAIS archive is free in implementing rights enforcement in whatever way it chooses. The only necessary restriction is to not introduce potential future barriers to the access by altering the raw Content Data Object; alterations due to encryption and watermarking of the raw data objects should only be applied when the content is finally presented to the user.Examples of PDI

Table 1 Examples of PDI components

Content Information Type	Reference Information	Provenance Information	Context Information	Fixity Information	Access Rights Information
Space Science Data	Object identifier Journal reference Mission, instrument, title, attribute set	Instrument description Principal Investigator Processing history Storage and handling history Sensor description Instrument Instrument mode Decommutation map Software interface specification Information Property Description	Calibration history Related data sets Mission Funding history	CRC Checksum Message Digest Reed-Solomon coding	Identification of the properly authorized Designated Community (Access Control) Permission grants for preservation and for distribution Pointers to Fixity Information and Provenance Information (e.g., digital signatures, and rights holders)

Content Information Type	**Reference Information**	**Provenance Information**	**Context Information**	**Fixity Information**	**Access Rights Information**
Digital Library Collections	Bibliographic description Persistent identifier	For scanned collections: metadata about the digitization process pointer to master version For born-digital publications: pointer to the digital original Metadata about the preservation process: pointers to earlier versions of the collection item change history Information Property Description	Pointers to related documents in original environment at the time of publication	Digital signature Checksum Authenticity indicator	Legal framework(s) Licensing offers Specifications for rights enforcement measures applied at dissemination time Permission grants for preservation and for distribution Information about watermarking applied at submission and preservation time Pointers to Fixity Information and Provenance Information (e.g., digital signatures, and rights holders)
Software Package	Name Author/ Originator Version number Serial number	Revision history Registration Copyright Information Property Description	Help file User guide Related software Language	Certificate Checksum Encryption CRC	Designated Community Legal framework(s) Licensing offers Specifications for rights enforcement measures applied at dissemination time Pointers to Fixity Information and Provenance Information (e.g., digital signatures, and rights holders)

2.4.4 Information Packaging

Some way is needed to connect together the various items OAIS identifies for preservation. In order to do this OAIS defines a general Information Package, which logically contains a general Information Object, as follows.

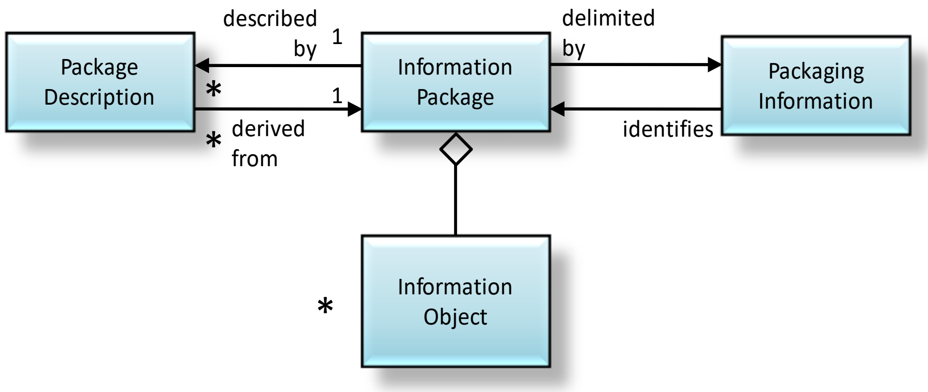


Figure 10 Information Package Contents and Its Associated Package Description and Packaging Information

Packaging Information is defined as the information that describes how the components of an Information Package are logically or physically bound together and how to identify and extract the components. It is a type of Information Object.

For example, all the components could be in a compressed file where the Information Object is in a directory called “InformationObject” which contains a Data Object and a subdirectory which in turn contains the Representation Information.. The Packaging Information would then be “This is a compressed file, compressed using specific compression algorithm ZZZZ, and the various components may be found, once the file is uncompressed, in directories “InformationObject” etc. Other examples will be discussed below.

Alternatively, there may simply be a text file where on each line starts with the name of the component e.g., “InformationObject” followed by a URI which points to the actual component. The packaging Information would then be “the bits which make up the Information Package can be understood as a text file with ASCII-7 encoding, and the Semantic Representation Information is pointed to by the URI labelled as SemanticRepInfo (for example).

Package Description is the information intended for use by Access Aids i.e., to discover the contents. It is a type of Information Object. In the example of the ZIP file above the Package Description could be a description of the contents of the Information Package.

OAIS defines three types of Information Package:

2.4.4.1 Types of Information Packages

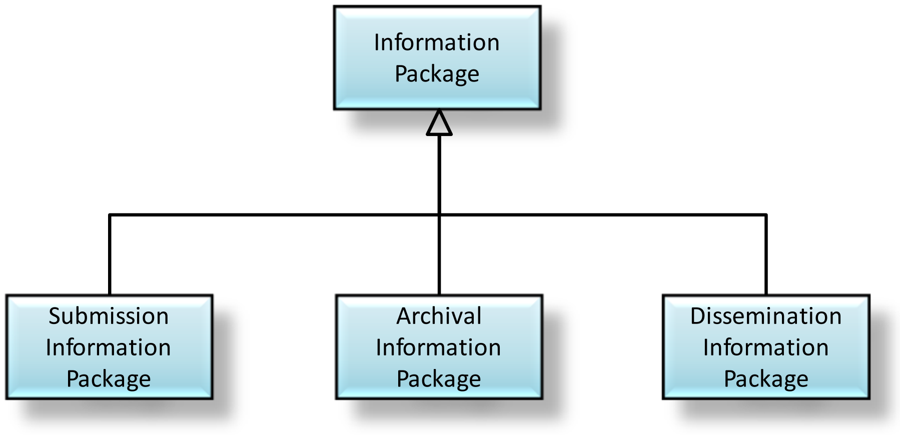

Information Package
Submission Information Package
Archival Information Package
Dissemination Information Package

Figure 11 Information Package Taxonomy

The definitions of Submission Information Package (SIP) and Dissemination Information Package (DIPt) essentially provide no additional details; they are defined as a convenience for the incoming and outgoing packages to and from the repository.

The Archival Information Package, on the other hand, is very specific.

2.4.4.2 Archival Information Package

The Archival Information Package is defined as a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite, Long-Term Preservation of a designated Information Object. In OAIS terms this means that it is an Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

The Information Object is then:

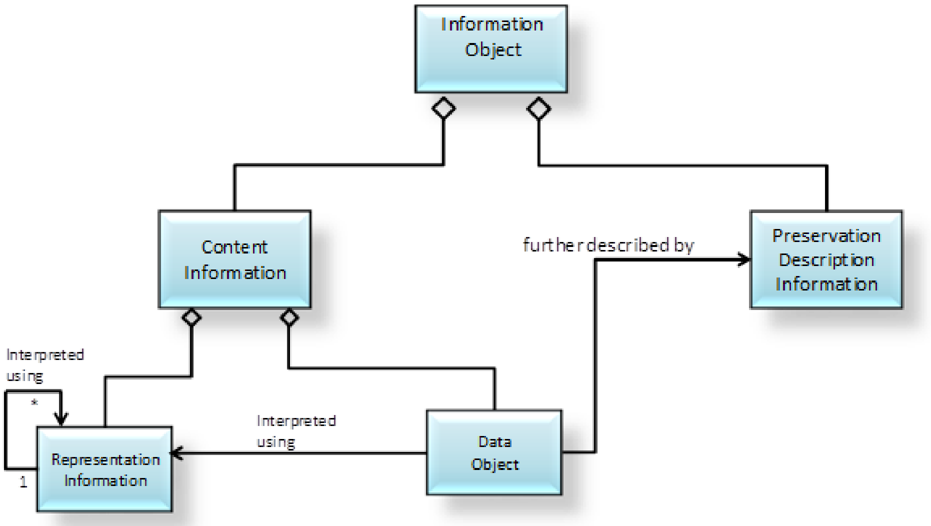


Figure 12 Example of an Information Object Made Up of Content Information and PDI

Expanding the figure to show all the components we get:

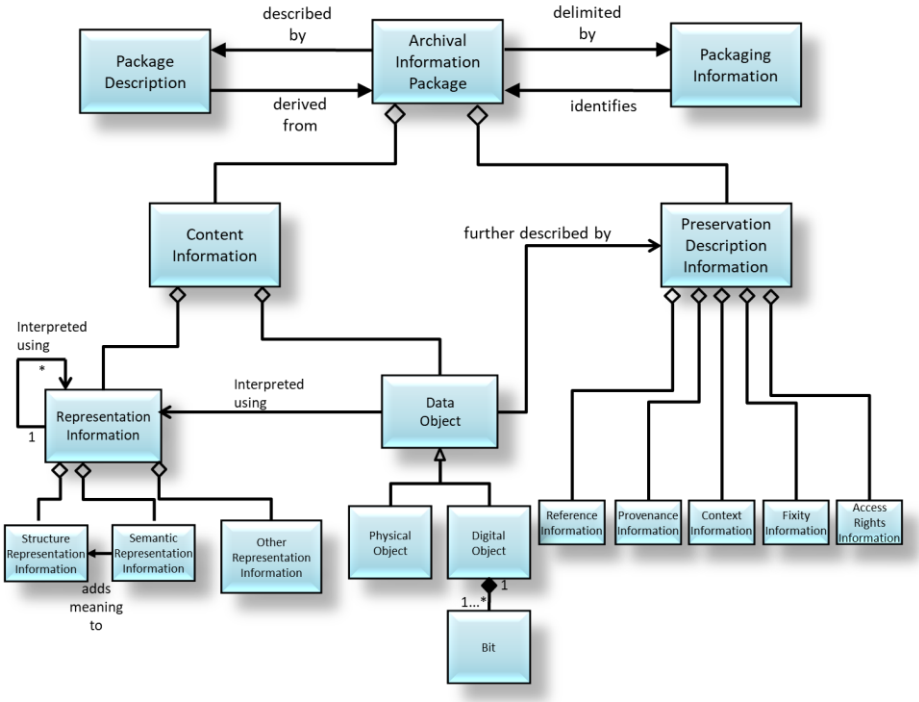


Figure 13 Archival Information Package (Detailed View) and Its Associated Package Description and Packaging Information

The AIP is a logical package – as are all the Information Packages, so could be just a collection of pointers. Of course, as time passes then Provenance Information will be added, for example as the Data Object is moved from one storage location to another, and Representation Information may be added as the Knowledge Base of the Designated Community changes.

If the AIP is all a single file, for example everything in a ZIP file, then the way to add information, for example by ensuring the AIP has some pointers to, in this case to the extra Provenance and Representation Information.

On the other hand, if the AIP is a text file pointing to the various components then as those components increment then the logical AIP is automatically incremented.

The reason an AIP is important is that it shows that the archive knows where all the components are, and could, if required, collect everything that is needed to preserve the Content Information and hand it over to another archive.

2.5 OAIS Functional Model

OAIS defines a number of functional entities and information flows, in order to provide useful terminology when describing an archive. One often sees the following Functional Model as representing the “OAIS Model”, but this is misleading.

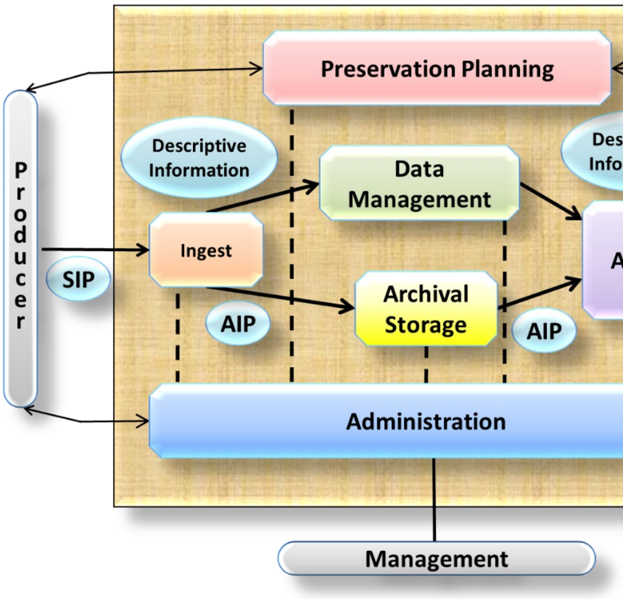


Figure 14 OAIS Functional Model

Consumer is defined as: The role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in whatever level of detail is allowed. In addition to the normally expected entities outside the OAIS, this can also include other OAISes, as well as internal OAIS persons or systems.

Producer is defined as: The role played by those persons or client systems that provide the information to be preserved. This can include internal or external OAIS persons or systems.

Management is defined as: The role played by those who set overall OAIS policy as one component in a broader policy domain, for example as part of a larger organization.

Note that these are all **roles**, which can be played by people or systems, and so a single person or system may play several roles at the same time.

Some go to extremes:

Figure 15 OAIS Functional Model tattoo[38]

[38] Courtesy of Flickr user wlef70 / Creative Commons Licensed - See http://britishlibrary.typepad.co.uk/collectioncare/digital-preservation/#sthash.tlUOZCc0.dpuf

The false belief that the Functional Model represents all of OAIS has led some to remark that a chicken with its head cut off could be OAIS compliant.

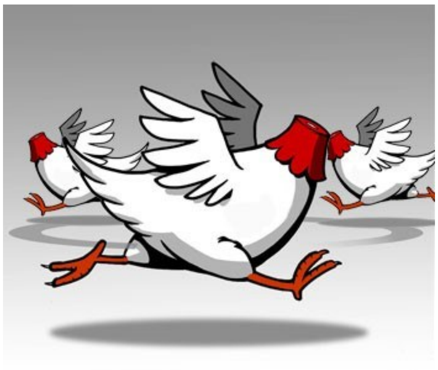


Figure 16 Headless chicken

The next section shows how wrong that statement is.

More importantly ***this shows that some people who make a profession of digital preservation could not be bothered to read the OAIS standard and yet feel competent to make statements about it and digital preservation in general***.

The Functional Model is sometimes the only one shown when OAIS is discussed. Sometimes only this model is mentioned when talking about OAIS compliance. If you meet people who make this mistake then please pay no attention to anything they say about digital preservation because clearly they do not take their responsibilities seriously.

In fact, as we will see, when we look at OAIS conformance, the Functional Model is NOT mentioned.

Why is it included in OAIS? We, the authors, felt that it was a useful addition to the document and that it performs a VITAL job – namely to help define TEMINOLOGY and act as a CHECKLIST for repositories. The Functional Model and the associated detailed version shown in Figure 17 look rather like a design BUT it was not meant to be that. The useful terminology checklist of activities that an archive should probably perform is not exhaustive nor are they mandatory, but if a repository omits them then at least the repository staff should consider whether something is missing. OAIS does not specify how any of these are performed nor the way they may be combined.

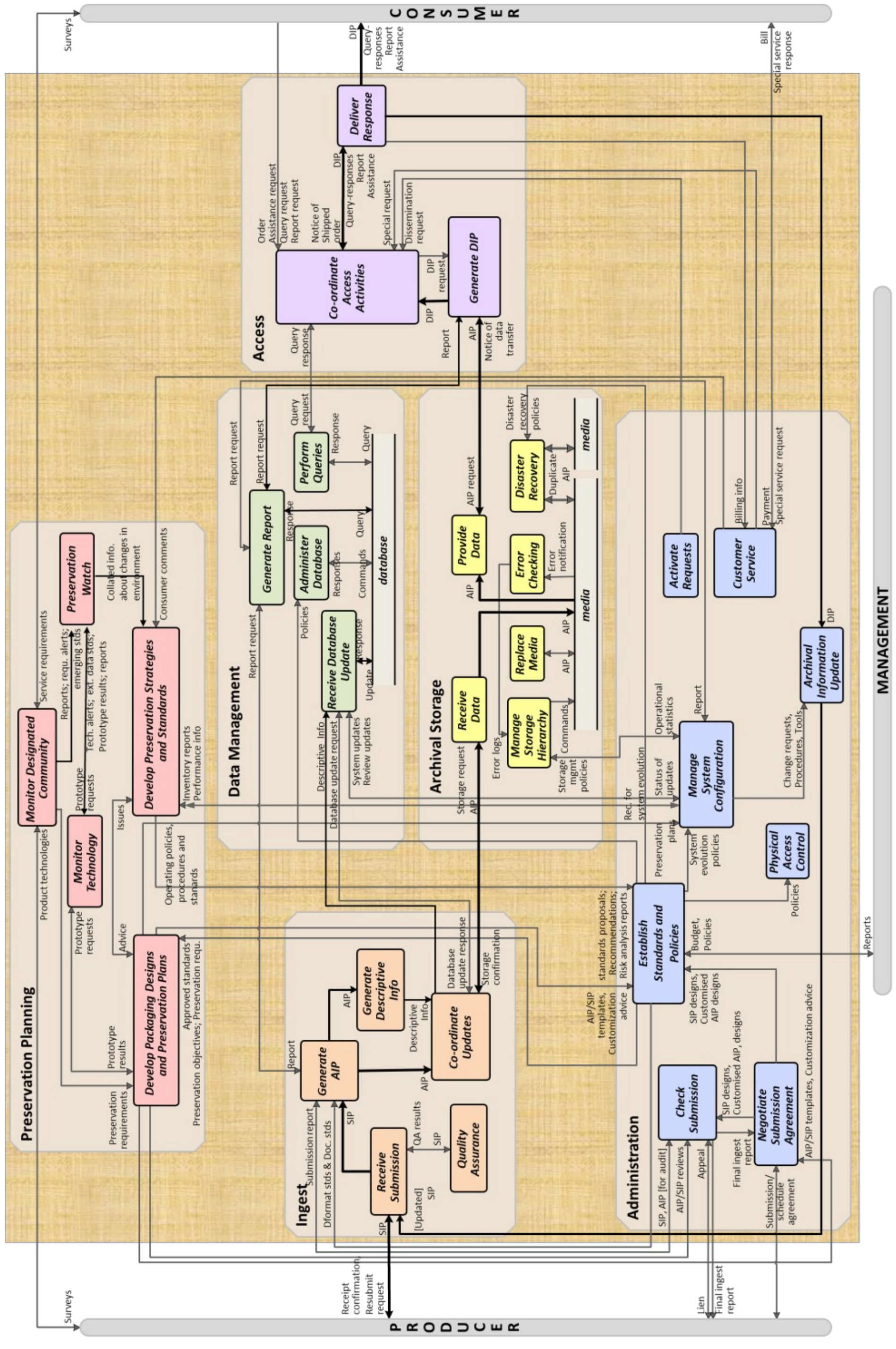


Figure 17 Detailed OAIS Functional Model Common Services assumed

Modern, distributed computing applications assume a number of supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, backup and directory services. Much excellent work has already been done in the area of open system environment reference models. Examples of such services include:

Operating system services provide the core services needed to operate and administer the application platform and provide an interface between application software and the platform. These services include the following:

- Kernel operations provide low-level services necessary to create and manage processes, execute programs, define and communicate signals, define and process system clock operations, manage files and directories, and control input-output processing to and from the external environment.
- Commands and utilities include mechanisms for operations at the user level, such as comparing, printing, and displaying file contents; editing files; pattern searching; evaluating expressions; logging messages; moving files between directories; sorting data; executing command scripts; and accessing environment information.
- Real-time extension includes the application and operating system interfaces needed to support those application domains requiring deterministic execution, processing, and responsiveness. The extension defines the applications interface to basic system services for input/output, file system access, and process management.
- System management includes capabilities to define and manage user resource allocation and access (i.e., what resources are managed, and the classes of access defined), configuration and performance management of devices, file systems, administrative processes (job accounting), queues, machine/platform profiles, authorization of resource usage, and system backup.
- Operating system security services specify the control of access to system data, functions, hardware, and software resources by users and user processes.

Network services provide the capabilities and mechanisms to support distributed applications requiring data access and applications interoperability in heterogeneous, networked environments. These services include the following:

- Data communication includes API and protocol specifications for reliable, transparent, end-to-end data transmission across communications networks.
- Transparent file access provides access to available files located anywhere in a heterogeneous network.
- Computer support provides support for interoperability with systems based on other operating systems, particularly computer operating systems, which may not be formally specified in a national or international standard.
- Remote Procedure Call services include specifications for extending the local procedure call to a distributed environment.
- Network security services include access, authentication, confidentiality, integrity, and non-repudiation controls and management of communications between senders and receivers of information in a network.

Security services provide capabilities and mechanisms to protect sensitive information and treatments in the information system. The appropriate level of protection is determined based upon the value of the information to the application end-users and the perception of threats to it. These services include the following:

- Identification/authentication service confirms the identities of requesters for use of information system resources. In addition, authentication can apply to providers of data. The authentication service may occur at the initiation of a session or during a session.
- Access control service prevents the unauthorized use of information system resources. This service also prevents the use of a resource in an unauthorized way. This service may be applied to various aspects of access to a resource (e.g., access to communications to the resource, the reading, writing, or deletion of an information/data resource, the execution of a processing resource) or to all accesses to a resource.

- Data integrity service ensures that data is not altered or destroyed in an unauthorized manner. This service applies to data in permanent data stores and to data in communications messages.
- Data confidentiality service ensures that data is not made available or disclosed to unauthorized individuals or computer processes. This service will be applied to devices that permit human interaction with the information system. In addition, this service will ensure that observation of usage patterns of communications resources will not be possible.
- Non-repudiation service ensures that entities engaging in an information exchange cannot deny being involved in it. This service may take one or both of two forms. First, the recipient of data is provided with proof of the origin of the data. This protects against any attempt by the sender to falsely deny sending the data or its contents. Second, the sender of data is provided with proof of delivery of data. This protects against any subsequent attempt by the recipient to falsely deny receiving the data or its contents.

2.6 Limitations of this chapter

This chapter summarises the key OAIS concepts which are important for preservation of all types of digitally encoded information. There is much more in the OAIS standard which is worth reading.

OAIS itself does not claim to cover all aspects of digital preservation, although what is does cover is fundamental. It also proposes a roadmap of follow-on standards, in particular the one needed in order to assess and certify archives. This will be covered in a later chapter.

3 Understanding what OAIS conformance actually means

In turns out that a great deal of the literature which talks about some kind of "OAIS conformance" has been written by people who do not seem to understand what OAIS conformance means. This chapter aims to ensure that the reader understands the concept clearly.

3.1 What OAIS says about conformance

The OAIS standard defines conformance as follows (the section numbers refer to those in the OAIS standard itself, not those in this book):

> *A conforming OAIS Archive implementation shall support, and be able to map to the components of, the model of information described in 2.3 and 4.3, which provides more formal definitions of the model using UML. The OAIS Reference Model does not define or require any particular method of implementation of these concepts.*
>
> *A conforming OAIS Archive shall fulfil the responsibilities listed in 3.2. Subsection 3.3 provides examples of the mechanisms that may be used to discharge the responsibilities identified in 3.2. These mechanisms are not required for conformance.*
>
> *A conformant OAIS Archive may provide additional services that are beyond those required of an OAIS.*
>
> *This reference model does not specify a design or an implementation. Actual implementations may group or break out functionality differently.*
>
> ...
>
> *To summarize, 1.6.2, 2.3, 3.2 and 4.3 are normative and form the basis of conformance to this Recommended Practice. All other sections are informative, but could aid the user in designing archives that conform to this Recommended Practice.*

We can look at these in more detail.

Some people erroneously talk about OAIS "compliance" rather than "conformance". The reason this is wrong is that "compliance" normally refers to something that this is a legal requirement, rather than something which must be followed to adhere to the standard.

3.2 OAIS Information Model

Conformance is divided into two segments, with the first requirement for a conformant OAIS archive being that it:

> *shall support, and be able to map to the components of, the model of information described in 2.3 and 4.3, which provides more formal definitions of the model using UML.*

The reason that the requirement is described in this way is that section 2.3 described the Information Model in very high-level terms, which are useful in terms of introducing the concepts but do not have enough detail for checking conformance. In the latest version of OAIS it was thought advisable to explicitly point out that the more detailed, formal, definition of the Information Model is in section 4.3, which contains detailed UML diagrams.

The point to note here is that an archive does not need to use the OAIS terminology in its documentation BUT if asked to point to, or provide copies of, Representation Information for a specific Data Object, then it should be able to do so, no matter what its internal terminology is.

Similarly, if an archive claims that it has Archival Information Packages (AIPs) then it must be able to identify, in whatever it calls its version of an AIP, all the various components which an OAIS AIP must have.

If it cannot do this then it cannot be conformant to OAIS.

3.3 OAIS Mandatory Responsibilities

The other requirement for conformance is that it must support the responsibilities in section 3.2 of the OAIS standard. These responsibilities may be explained as follows.

3.3.1 Negotiate for and accept appropriate information from information Producers.

WHY: The reason for this requirement is that many times in the past digital objects have essentially been dumped on an archive with little or no documentation about it, making them practically impossible to preserve. In order to help prevent this the archive should make an agreement with the Producer for the hand over not just of the digital objects but also the Representation Information and Preservation Description Information which includes, amongst other things, Provenance Information.

HOW: OAIS does not give a model for such an agreement, but the follow-on standards PAIMAS (ISO 20652) - see http://public.ccsds.org/publications/archive/651x0m1.pdf and PAIS (ISO 20104) - see http://public.ccsds.org/publications/archive/651x1b1.pdf provide some guidelines.

3.3.2 Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.

WHY: The issue here is that the archive needs physical as well as legal control over the information. The need for physical control is fairly obvious, for example to ensure that the bits are safe. Legal control is required because copyright and other legal restrictions, which may be different from one country to the next and may change over time, could otherwise limit the copying and migrations that the archive almost certainly will have to perform. While the lack of such legal control might not stop the archive performing such copying, nevertheless there is a risk that subsequent legal action may force the archive to stop and delete such copies or face financial penalties which could, at the extreme, cause the archive to cease operations.

HOW: When acquiring the Content Information from any other producer or entity, the OAIS should ensure that there is a legally valid transfer agreement that either

- transfers intellectual property rights to the OAIS,
- or clearly specifies the rights granted to the OAIS and any limitations imposed by the rightsholder(s).

The OAIS should ensure that its subsequent actions to preserve the information and make it available conform with these rights and limitations.

When the OAIS does not acquire the intellectual property rights, the agreement should specify what involvement the rightsholder(s) will have in preservation, management, or release of the information.

In most cases, it will be preferable for the OAIS to negotiate an agreement that specifies the rightsholder(s) requirements and authorizes the OAIS to act in accordance with those requirements without active involvement of the rightsholder(s) in individual cases.

3.3.3 Determine, either by itself or in conjunction with other parties, which entities should become the Designated Community. i.e., the communities that should be able to understand the information provided. Definition of the Designated Community includes a determination of their Knowledge Base.

WHY: As discussed earlier, it is essential for the archive to define the Designated Community for a data set in order for preservation to be tested. The definition of the Designated Community allows the archive to be clear about how much Representation Information is needed.

HOW: The Designated Community for a piece of digitally encoded information is not set in stone – it is a decision for the archive (possibly after consulting other stakeholders). It may reasonably be asked "What's to stop the archive making its life easy by defining the Designated Community which is easiest for it to satisfy?" It could for example just say, "The Designated Community is that set of people who understand these bits". The answer to the question may be understood by asking oneself the following: "Would I trust my digital objects to an archive which adopts such a definition of Designated Community?" It is to be hoped that it would be fairly self-evident that the use of such a definition would lead to a rapidly diminishing set of people who could understand the digital objects and therefore the archive could not really be said to be doing a good job. Therefore, depositors will, if they know that the archive uses such a definition, will not wish to entrust their valuable digital objects to such an archive. Thus, it is the "market" which keeps the archive honest. As will be clear when we discuss audit and certification, this definition(s) the archive adopts have to be made available. The question then arises from the point of view of the archive:" How should I define a Designated Community?" OAIS provides no explicit guidance on this point.

3.3.4 Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.

WHY: As discussed earlier the "Independently Understandable" aspect is to make it clear that a member of the Designated Community cannot simply pick up the phone and ask one of the people who created the digital objects for help. This is a practical consideration because such a phone call may be possible when the data is deposited, but certainly will not be possible in 200 (or even 20) years' time. This is not a one-off responsibility. It is one which must continue into the future as the Knowledge Base of the Designated Community changes.

HOW: The archive must have adequate Representation Information in order to satisfy this responsibility. This means that it must be able to create, or have access to, Representation Information, and it must be able to determine how much is needed. These key requirements require the kinds of tools which are discussed in subsequent chapters; Chap. 7 describes many techniques for creating Representation Information and describes where each technique is applicable.

3.3.5 Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.

WHY: This responsibility states the fairly obvious point that the archive should look after the information in the basic ways e.g., against floods and theft. The demise of the archive deserves special consideration. Although many archives act as if they will always exist with adequate funding, this particular responsibility points out that such an assumption must be questioned. In addition, of course the archive should not be able to delete its holdings on a whim. Many might take the view that deletions should never be allowed, however others insist that deletions are a natural stage in the life of the data. The wording of this responsibility allows the archive to make such deletions but only under (its own) strictly defined circumstances.

HOW: Backup policies and security procedures should take care of the "reasonable contingencies" as long as they are adequate, including

- guarding against Natural Disasters and Theft
- deleting information **only** according to Established Policy

While it is not possible to guard against the demise of the archive, for example if funding dries-up, nevertheless it is possible to make plans to safeguard the digital objects by making agreements with other archives. Such agreements would provide a commitment by the second archive to take over the preservation of the digital objects. Of course, since one cannot be sure which other archives will continue to exist, an archive may make agreements with several other archives, and perhaps different archives may agree to take different subsets of the holdings. To move information to another archive requires that complete AIPs are copied.

3.3.6 Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Content Information with evidence supporting its Authenticity.

WHY: There are two parts to this responsibility. The first is that the digitally encoded information has to be made available, at least to the Designated Community. The second part contains a new requirement which is introduced here because we are talking not about understandability, which many other responsibilities cover, but about access. The key question concerns how a user can have confidence that the digital object which the archive provides to him/her is authentic i.e., what it is claimed to be. Section 4 contains a detailed discussion of Authenticity. The phrase "copies of, or as traceable to" means that the archive may keep the original bits and send a copy to the user, or it may have performed various operations such as sending only a sub-set of the original or carried out preservation activities, such as transformation, which change the bit sequences, but will have to maintain appropriate evidence.

HOW: The way in which digital objects are made available to any users are many and varied. In fact, access is the user-facing part of the archive where it can make its mark and an immediate impression on users and potential users. OAIS has very little to say about the types of access which may be provided. Dark Archives are those which hold digital objects but do not make them accessible – at least not for some period or until some pre-determined trigger. These archives can still be preserving the understandability and usability of the digital objects for a Designated Community but do not, during that "dark" period, allow even the Designated Community to access them. During that "dark" period it would not be possible, without special access being granted, to verify the preservation of those digital objects.

3.4 Limitations of this chapter

This chapter summarises the aspects of OAIS related to conformance. There is much more in the OAIS standard which is worth reading.

The limitation is that it does not have enough detail to be the basis of a full audit of a repository – nor was it meant to be. On the other hand, it is possible to judge whether a repository does **NOT** conform to OAIS for example if it cannot identify all the elements of the AIP or cannot support all the Mandatory Responsibilities.

As mentioned above, it is surprising that many people who claim to know about OAIS have clearly not read the very short section (section 1.4 in OAIS) on conformance, named rather helpfully "CONFORMANCE".

4 Explanation of updates to OAIS version 2012

The definitions in this book have been taken from the updated version of OAIS, whereas many readers may be familiar with the 2012 version of OAIS. This chapter describes and explains the changes made[39].

4.1 Updates to OAIS Concepts

4.1.1 Representation Information

One of the key OAIS concepts is Representation Information, which, when combined with a Data Object, produces an Information Object. The question as to how much Representation Information is needed is determined by the definition of the Designated Community and its Knowledge Base.

The amount of Representation Information will change over time as the Knowledge Base of the Designated Community changes. The OAIS needs to ensure that it has Long Term access to all the relevant Representation Information. A choice must be made whether the OAIS collects all the relevant Representation Information itself or references the existence of the Representation Information in another trusted OAIS Archive. That choice is an implementation and organization decision.

The updates make it clear that **in special cases** the initial amount of Representation Information required may be very minimal. For example, *for a specific Data Object and a specific Designated Community, the Knowledge Base of the Designated Community is adequate for its members to understand or use the Data Object. In such cases the Representation Information could be the statement that no additional Representation Information is needed for that specific Designated Community at that particular time*, but further Representation Information may need to be collected in future. The revised text goes on to say, …"*any Representation Information that can be gathered at ingest should be included since it will likely be costlier to rediscover and add it at a later time*."

4.1.2 Preservation Description Information (PDI)

In the versions of OAIS up to now the components of PDI, namely Provenance Information, Reference Information, Fixity Information, Access Rights Information and Context Information, referred to the Content Information, i.e., the Content Data Object plus its Representation Information. Although these are a consistent and useful set of definitions, it does cause some problems in terms of potential implementations. Consider the case where one deals with a distributed network of Representation Information, which changes with the

[39] Extracted from Giaretta, D. et al, OAIS Version 3 Draft Updates, iPRES 2019, 2019, https://ipres2019.org/static/pdf/iPres2019_paper_51.pdf

Designated Community's Knowledge Base. A change in some part of the Representation Information network would mean that all the elements of the PDI would change.

The update concerning PDI is that all the components of PDI would now refer to the Content Data Object rather than Content Information.

There are several reasons for this change. The consensus was that for most, perhaps all, repositories, the PDI components do refer to the Content Data Objects. For example, the Fixity Information is often essentially a digital digest of the Content Data Object. This focus on Data Objects would also make audits of repositories more practical since the auditor can perform checks on specific Content Data Objects. Of course, even the Content Data Object may be complex, for example consisting of many files, but at least changes in the Knowledge Base of the Designated Community does not cause it to change.

A related point considered by the group was that, for example, the Representation Information should have Fixity also. To clarify this, point the following note was added to emphasize the fact that, from the very first version of OAIS, the Information Model applies to every one of the things which are called "Information", including, for example, Representation Information and Provenance Information.

Defining PDI (as well as its components - Provenance Information, Context Information, Reference Information, Fixity Information, and Access Rights Information) as relevant to the Content Data Object does not mean that those concerns are any less important for other data objects or at other levels, for example, it is important to apply reference, fixity, provenance, context and access rights to Representation Information, or to any other information the Archive is preserving. Definition of these terms as relevant to the Content Data Object is simply to ease discussion of these concepts at the Content Data Object level.

In other words when one is talking about, for example, Representation Information as the target of preservation, then one can regard it as Content Information in its own right, as well as being part of another instance of Content Information. To some readers this may seem a strange way to describe things, but it is similar to what should be the familiar arrow in the OAIS Information Model which "loops back" from Representation Information back to itself.

4.1.3 Preservation Objectives

Usability has played a central role in defining preservation. However, there was a feeling that the meaning of usability needed to be clearer, and more testable. To this end the concept of a "Preservation Objective" has been introduced and defined as a *specific achievable aim which can be carried out using the Information Object*.

Preservation Objectives can then be used in the definition of other terms including:

- *Representation Information: The information that maps a Data Object into more meaningful concepts. so that the Data Object may be understood in ways exemplified by Preservation Objectives.*

- *Independently Understandable: A characteristic of information that is sufficiently complete to allow it to be understood by the Designated Community, as exemplified by the associated Preservation Objectives, without having to resort to special resources not widely available, including named individuals*

Preservation Objectives are intended to allow the repository to make it possible to test and demonstrate whether the information actually is Independently Understandable by members of the Designated Community now and into the future.

Examples of Preservation Objectives are provided in the updated OAIS:

- *The ability to render documents, images, videos, or sounds in a way which is sufficiently similar to the original. This could be checked by verifying that, for example, the document is readable,* *or the image is viewable. An analysis of the colours could also be compared. A spectral analysis could be performed on the sounds and compared with that of the original.*

- *The ability to process a dataset and generate the data products expected. This could be checked by comparing with something generated earlier, for example on Ingest.*
- *The ability to understand a dataset and use it in analysis tools to generate results, for example the density of electrons in the upper atmosphere or the structure of a molecule, given certain measurements. These could be compared with results generated earlier.*
- *The ability to re-perform an artistic performance. This could be compared with a recording of a previous performance.*

4.2 Update to the OAIS Functional Model

There have been many small clarifications made to the various text and diagrams which make up the Functional Model, introducing unambiguous shapes for diagram entities; MOIMS-DAI hopes that CCSDS/ISO will allow the publication of the new version to include the colours which give visual clues as to the grouping of the boxes.

The one area where an extra function has been added is to the Preservation Planning Functional Entity.

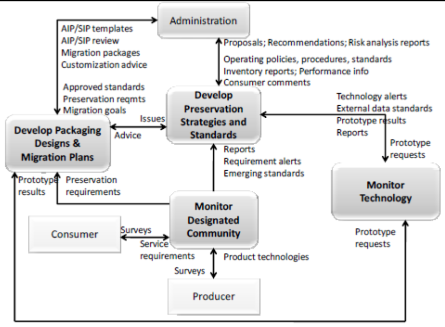
<table>
<tr>
<td>


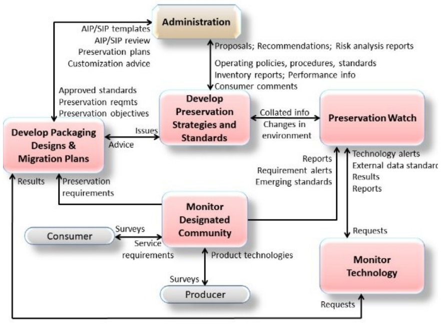
Figure 18 2012 version of Preservation Planning
</td>
<td>


Figure 19 Updated Preservation Planning Functional Entity
</td>
</tr>
</table>

The additional function is the already widely used “Preservation Watch.” This is described in the update as follows:

The Preservation Watch function is the role of collating preservation related information from a variety of internal and external entities. The Preservation Watch function also brings in reports, requirement alerts and emerging standards from the Monitor Designated Community function and technology alerts, external data standards, results and reports from the Monitor Technology function. Changes in the environment of the Archive (financial, political, and environmental) can be part of the Preservation Watch function.

Previously, Preservation Watch functionality was primarily located within the Develop Preservation Strategies and Standards.

4.3 Updates to the OAIS Information Model

The major updates to the Information Model carry forward the changes which have been described in section 4.1. These are summarized in the following diagram where the PDI connects to the Data Object rather than the Content Information:

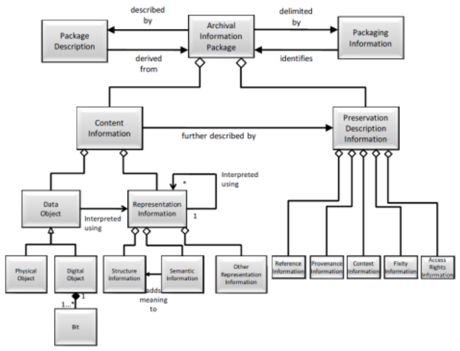


Figure 20 2012 version of AIP

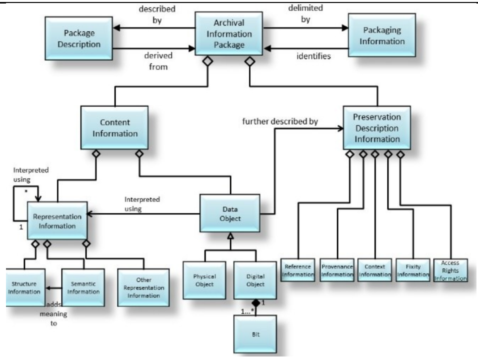


Figure 21 Updated AIP diagram

4.3.1 Updates to Information Package Definition

An Archival Information Package is the most detailed example of an Information Package, one which must contain Content Information as well as PDI. However, SIPs and DIPs do not need to contain any of these components.

The update to the general OAIS Information Package, shown next:

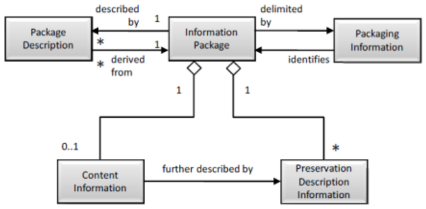


Figure 22 2012 version of Information Package

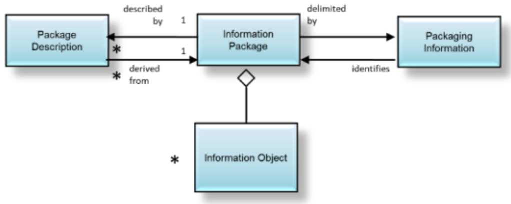


Figure 23 Updated Information Package

This change makes it clear that SIPs and DIPs can be defined in a much more flexible way. Note that this does not require any changes to the definition of the AIP because, as illustrated in Figure 24, the combination of Content Information and PDI can be regarded as a single, albeit complex, Object, made up of multiple Information Objects.

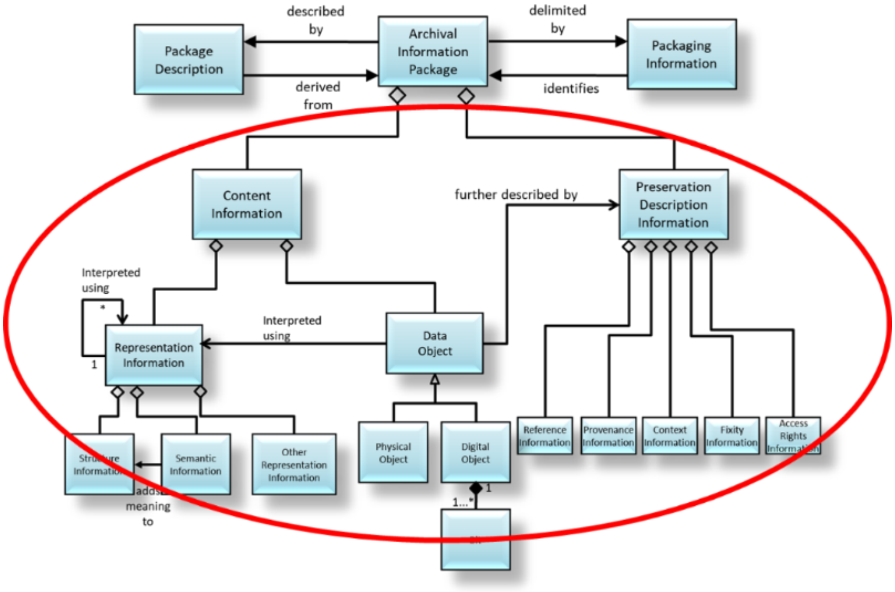


Figure 24 The combined Information Object in an AIP

4.4 Updates to Preservation Perspectives

Major changes have been made to the section of OAIS which describes practices that have been, or might be, used to preserve digital information and to preserve access services to digital information.

Up to now, essentially the only preservation practice which has been explicitly described has been Migration and Preservation of Access, e.g., Emulation. However, clearly the OAIS mandatory responsibilities require that there be adequate Representation Information, and that the OAIS should preserve information against all reasonable contingencies, including the demise of the Archive.

The changes in the new draft now include explicitly that the Content Data Object being preserved may be

(1) kept by the Archive but may be changed or

(2) kept by the Archive unchanged or

(3) not kept by the Archive, but instead be handed on to another Archive

Each of these three imply the following:

In case (1) the Archive may Transform the Content Data Object

In case (2) the Archive may add Representation Information to ensure the Content Information is Independently Understandable

In case (3) the Archive may hand over the AIP which contains the Content Data Object

This change makes the text as a whole more consistent and clearer.

4.5 Updates to Archive Interoperability

A major change to the discussion of various possible types of archive interactions is the way in which the distribution of OAIS functionality may be described. Such a distribution of functionality could be motivated, for example, by cost reduction or the availability of a comprehensive functionality offer. These descriptions should allow archives to be described more accurately and make it even clearer that an OAIS has never been required to be a monolithic organisation.

The text describes some possible categories (not an exhaustive or mutually exclusive list) of Archive associations. The first set of three categories has successively higher degrees of organizational interaction:

- *Independent: Archives motivated by local concerns with no management or technical interaction among them.*
- *Cooperating: Archives with potential common Producers, common submission standards, and common dissemination standards, but no common Finding Aids.*
- *Federated: Archives with both a Local Community (i.e., the original Designated Community served by the Archive) and a Global Community (i.e., an extended Designated Community) which has interests in the holdings of several OAIS Archives and has influenced those Archives to provide access to their holdings via one or more common Finding Aids.*

Another set of categories, somewhat orthogonal to the previous set, differentiates according to how internal Archive functions and functional areas are addressed and by styles of resource sharing.

- *All In-house: Archives that perform all archival functions in-house.*
- *Shared resources: Archives that have entered into agreements with other organizations to share resources, perhaps to reduce cost. This requires various standards internal to the Archive (such as ingest-storage and access-storage interface standards) but does not alter the user community's view of the Archive.*

- *Distributed: Archives that have distributed the OAIS functionality either geographically or organizationally. Different levels, forms and organization of the distribution are possible. In every case, the Archive is required to oversee and manage the Archive's use of the distributed functions, but does not alter the user community's view of the Archive*

An important classification of distribution is where the supporting organizations, which supply the required functionality, are themselves each an OAIS. One can describe the arrangement as a primary OAIS using one or more supporting OAIS for specific services. In such a case, each supporting OAIS, as well as the primary OAIS must fulfil all requirements for OAIS conformance, namely the Mandatory Responsibilities and support for the Information Model. Therefore, service level agreements are required to guarantee proper implementation of the functionality distribution. Particularly, the primary OAIS must monitor that the supporting OAIS is meeting its service agreement. The conformance of each supporting OAIS may be used as a piece of evidence.

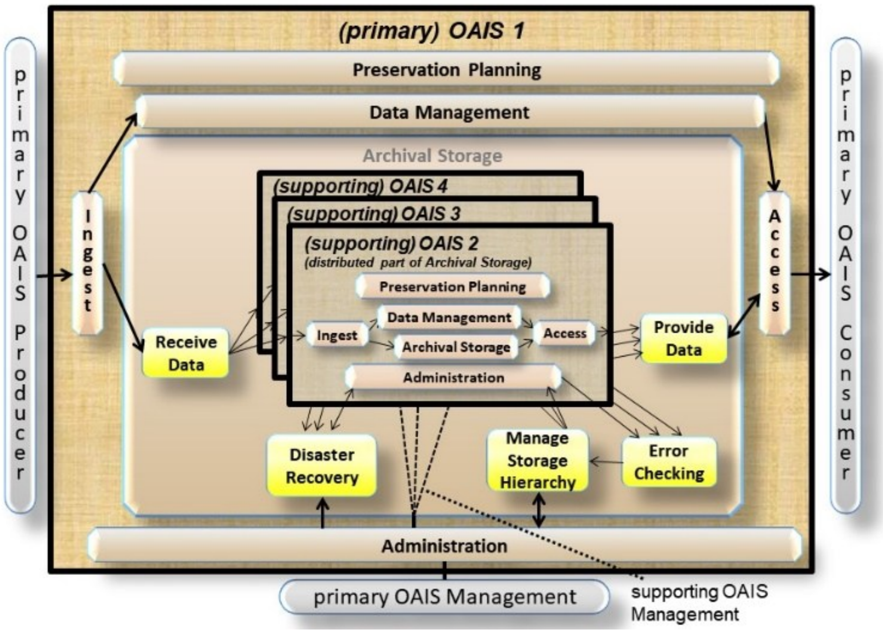


Figure 25 Primary/Supporting OAIS distributed functionality

The term 'Outer OAIS-Inner OAIS' has been used in the literature[40]. This usage is consistent with the "Outer OAIS" being the primary OAIS and the "Inner OAIS" being the supporting OAIS in cases where the "Outer OAIS" and "Inner OAIS" are each totally conformant to OAIS requirements. To exemplify the use of distributed functionality with supporting (inner) OAISes the Figure 25 shows how a set of supporting OAISes complete the functionality of the primary OAIS Archival Storage.

4.6 Limitations of this chapter

This chapter has presented a factual summary of the changes in OAIS.

[40] [3] "Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS – Inner OAIS (OO-IO) Model" by Eld Zierau and Nancy McGovern. In Proceedings of the 11th International Conference on Preservation of Digital Objects (iPRES) 2014, pp. 209-218, available at http://www.ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf

5 Authenticity

As we have seen in previous chapters, the bits may be "written" in many different ways and are likely to be "re-written" many times. If we are preserving some digitally encoded information it is therefore important to be sure that what we are preserving is what we think it is. This chapter describes the ways in which this can be achieved.

5.1 What is authenticity?

authenticity (plural authenticities)

1. *The quality of being genuine or not corrupted from the original.*

   I hereby certify that this is an authentic copy.

2. *Truthfulness of origins, attributions, commitments, sincerity, and intentions.*

   The painting was not authentic after all; it was just a copy.

3. *(obsolete) The quality of being authentic (of established authority).*

*(**Wiktionary definition** downloaded 11 Aug 2022)*

Authenticity is a fundamental issue for the long-term preservation of digital objects: the relevance of authenticity as a preliminary and central requirement has been investigated by many international projects. Some focused on long-term preservation of authentic digital records in the e-government environment, and in scientific and cultural domains.

Much has been written about Authenticity. However, in order to create tools which can be relied upon, and which are practical we must achieve the following:

- build on the excellent work which has already been carried out
  - previous work has for the most part focussed on what we have referred to as Rendered Digital Objects therefore we must ensure that we can deal with the variety of other types of objects.
- convert these rather abstract ideas into something which is widely applicable but also practical and implementable
- show some practical examples, from real archives, using a practical tool.

Therefore, this chapter first discusses the previous work on Authenticity, including the definitions from OAIS, and introduces a number of basic concepts. From these concepts we build up a conceptual model. The concept of Significant Properties has been much discussed, especially with respect to Rendered Digital Objects. We show how this can be extended to Non-rendered Digital Objects, and how it fits into the work on Authenticity. Finally, we apply these concepts and models to a number of digital objects from real archives, using a tool based on the conceptual model.

5.2 Background to Authenticity

Authenticity is a key concept in digital preservation, and some would argue that it is **the** pre-eminent concept, in that unless one can show that the data object is, in some provable sense, what was originally deposited, then one cannot prove that digital preservation has been successful.

On the other hand, OAIS defines preservation in terms of understandability and usability as well as authenticity; it therefore provides a view in which Representation Information and Authenticity are equal partners.

It is worth noting the distinction made by InterPARES[41], although specifically referring to *records*, between verification and maintenance of authenticity. Some would argue that everything is a *record,* but this point will not be discussed here.

Verification of authenticity is "the act or process of establishing a correspondence between known facts about the record and the various contexts in which it has been created and maintained, and the proposed fact of the record's authenticity"[42].

Maintenance of authenticity is related to records which "have been presumed or verified authentic in the appraisal process and have been transferred from the creator to the preserver".

This chapter discusses both the **maintenance** of authenticity i.e., providing a continuing chain of evidence about the custodianship and treatment of the information as well as the **verification** of authenticity in so far as a Consumer must be able to make that judgement.

5.2.1 Links to previous literature

A separate position paper[43] reviews the InterPARES authenticity work in more detail; the main conclusions from that paper are included in this chapter. However, it is worth mentioning two concepts which are regarded in the literature as crucial, namely ***integrity*** and ***identity*** of digital resources; authenticity is regarded as being established by assessing the integrity and the identity of the resource.

The ***integrity*** of a resource refers to its wholeness. A resource has ***integrity*** when it is complete and uncorrupted in all its essential respects..

The ***identity*** of a resource, from this point of view, has a very wide meaning, beyond its unique designation and/or identification. Identity refers to *the whole* of the characteristics of a resource that uniquely identify it and distinguish it from any other resource. In addition to its internal conceptual structure, it refers to its general context (e.g., legal, technological). From this point of view, identity is strongly related to PDI: Context, Provenance, Fixity, Reference, and Access Rights Information, as defined in OAIS, help to understand the environment of a resource.

This information has to be gathered, maintained, and interpreted together – as far as possible – as a set of relationships defining the resource itself: a resource is not an isolated entity with defined borders and autonomous life, it is not just a single object; a resource is an object *in the context*, it is both the object itself and the relationships that provide complete meaning to it. These relationships change over time, so we need not only to understand them and make them explicit but also to document them to have a complete history of the resource: we cannot omit it without also losing a bit of the identity of the resource, with consequences on its authenticity.

[41] InterPARES Project http://www.interpares.org

[42] InterPARES Project, Authenticity Task Force, Requirements for Assessing and Maintaining the Authenticity of Electronic Records, March 2002 (http://www.interpares.org)

[43] International Study on the Impact of Copyright Law on Digital Preservation (2008) http://www.digitalpreservation.gov/library/resources/pubs/docs/digital_preservation_final_report2008.pdf

5.3 OAIS Definition of authenticity

It must be admitted that in the original version of OAIS authenticity was not dealt with very well; however, the OAIS updates have significantly improved the situation and we use the definitions from that update.

OAIS defines **Authenticity** as:

> ***the degree to which a person (or system) may regard an object as what it is purported to be. The degree of Authenticity is judged on the basis of evidence.***

We have the associated definition:

Provenance Information *The information that documents the history of the Content Data Object. This information tells the origin or source of the Content Data Object, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity.*

The concept of authenticity as defined by OAIS requires a detailed analysis on the basis of the rich literature in the sector and of the main outputs of research projects like InterPARES.

There are some basic assumptions to be considered before entering into a detailed analysis. First of all, authenticity cannot be evaluated by means of a YES/NO flag telling us whether a document is authentic or not. In other words, there are degrees of confidence about the authenticity of the digital resources: certainty about authenticity is a goal, but one which is unlikely to be fully met.

In the case of physical objects such as a parchment, one can look at physical composition, confirm the age with carbon dating, and compare to similar documents. The ***identity*** and ***integrity*** concepts can fairly obviously be applied to such physical objects which cannot be readily copied of changed without leaving some trace. What is so different with digital objects? What process of evaluation and what sort of tools must be developed?

One fundamental issue arises from the basic points about why digital objects are different, namely that one cannot really ensure the ability to maintain the original bits or even to provide methods for easily evaluating whether they are the original. At the very least one has to copy the bits from one medium to another. How can we be sure that the copy was done correctly? Here we want to guard against both accidental changes as well as changes made on purpose, perhaps for nefarious purposes.

Other issues arise from the basic preservation strategies. Adding Representation Information, maintaining access and emulation do not require any changes to the bit sequences of the digital object, nor do the types of migration described as refreshment of replication. Repackaging requires changes in the packaging but not of the digital object of interest. In all these cases we can use, for example, digests as the evidence about the lack of change in the bit sequences.

Only the ***Transformation*** preservation technique implies a change in the digital object and therefore the use of digests will not apply. As an example of Transformation, consider a WordStar version of a document which may be converted into Word 2007 format in order to ensure that this (Rendered) Digital Object can continue to be rendered. Is it an authentic version, and what does that means in this context?

Looking in more detail at the use of digests, one compares the digest of the original to the copy and then one can be fairly, but not completely, certain things are as they should be. It would be alright if one has access to the original and could calculate the digest oneself, but this is rarely the case. So where does the digest come from, and how can we be sure it is indeed the digest i.e., that it is what it purports to be? This sounds like a familiar question - we have come round in a circle! This is of course another example of recursion.

So how does this help us with the Authenticity of digital objects? It seems that the key points are that

- we need various types of evidence. The questions then become:
  - what evidence must be collected – are there procedures to follow?
  - from whom?, and
  - how are we going to be sure the evidence has not been altered?
- we need to be able to trace the evidence back to someone or something. The follow-on questions are then:
  - can we identify the person?
  - can we be sure that that person supplied the evidence?
  - how can we be sure that the person identified is the person claimed, and that this information has not been altered?
  - is the person trustable?
- we need to have a view on how this evidence can be evaluated

These ideas have clear links to the OAIS definitions which will be expanded below.

As was mentioned, we focus on the **maintenance** of (evidence about) authenticity. Defining and assessing authenticity in the repository (or more formally – in the custodial environment) – on which we focus here – are complex tasks and imply a number of theoretical and operational/technical activities. These include a clear definition of roles involved, coherent development of recommendations and policies for building trusted repositories, and precise identification of each component of the custodial function.

Thus, it is crucial to define the key conceptual elements that provide the foundation for such a complex framework. Specifically, we need to define how, and on what basis authenticity has to be managed in the digital preservation processes in order to ensure the trustworthiness of digital objects. In order to do this, we need a more formal model about authenticity.

The authenticity of digital resources is threatened whenever they are exchanged between users, systems or applications, or any time technological obsolescence requires an updating or replacing of the hardware or software used to store, process, or communicate them. Therefore, the preserver's inference of the authenticity of digital resources must be supported by evidence provided in association with the resources through its *documentation* (recursion again!), by tracing the history of its various migration and treatments, which have occurred over time. Evidence is also needed to prove that the digital resources have been maintained using technologies and administrative procedures that either guarantee their continuing *identity* and *integrity* or at least minimize risks of change from the time the resources were first set aside to the point at which they are subsequently accessed.

Let's go back a step and see what we can learn from a more familiar case. How do you prove that you are who you say you are? For people with whom you grew up there is probably no problem - they have seen you every day from a small baby to a grown up and know you are the same person (ignoring any philosophical digressions) despite all the changes that have happened to you physically. To prove who you are when you enter a new country you would present your passport - why is that accepted? It is a physical object (a little booklet) that can be examined and checked against some known standard document. But why is that accepted? The assumption is that it is backed by your government, but how does the government which issued it know that the person represented there is in fact you? In the UK, and perhaps elsewhere, one has to fill in a form and provide a photograph, both of which are signed by someone known, and probably trusted, in the community such as a doctor or local politician. How does the state know that that person is indeed who he/she is indeed the person who claims to be that doctor or local politician? There may be other information that the government has to cross check, but it could just rely on going back to the community and check - because they are known in the community in some way.

So how does this help us with the Authenticity of digital objects? It seems that a key point is that we need to be able to trace back to someone who is, in some way, trusted. Also, there is some type of evidence collected. The questions then become: what evidence must be collected, from whom, and how are we going to be sure the evidence itself is true? These ideas will be expanded below.

5.3.1 Transformational Information Properties

One of the ways to preserve digitally encoded information when circumstances change is to Transform the Data Object. However, this causes problems in terms of the Authenticity of the information. Clearly we cannot rely on the bits being the same because the object has been Transformed. What can be done? What evidence can be provided to support claims of authenticity. Previous literature has spoken in terms of using Significant Properties, however it became clear[44] that the definitions involved were not sufficiently clear, nor sufficiently general. The OAIS working group addressed these issues by defining a sequence of terms which could be combined to build the concepts required. The idea is that, as we have seen in other sections, there are many things in the bits of a digital object which **may not** be of importance to us. Therefore, we need a way to capture what **is** important to us. To do this, OAIS provides the following terminology.

Transformational Information Property (TIP):

An Information Property, the preservation of the value of which is regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately preserved information content. This could be important as contributing to evidence about Authenticity. Such an Information Property is dependent upon specific Representation Information, including Semantic Representation Information, to denote how it is encoded and what it means.

The term **Information Property** is defined as

> **A part of the information content of a Content Information object that is highlighted for a particular purpose.**

This is quite general and could be anything about the information being preserved. The Information Property Description provides a description of the Information Property

In order to explain why the TIP mentions Non-Reversible Transformation we just need to think of a Transformation is Reversible, defined as:

> **A Transformation in which the new representation defines a set (or a subset) of resulting entities that are equivalent to the resulting entities defined by the original representation. This means that there is a one-to-one mapping back to the original representation and its set of base entities.**

In such a case can simply reverse it and get back to the original object and that would provide the evidence of Authenticity required.

[44] Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G., & Sawyer, D. (2009). Significant Properties, Authenticity, Provenance, Representation Information and OAIS Information. *UC Office of the President: California Digital Library*. Retrieved from https://escholarship.org/uc/item/0wf3j9cw

Therefore, we need to define one or more Transformational Information Properties. What should they be? OAIS says:

The Producer may provide, or the Archive may itself define, Information Property Descriptions of Information Properties which should be maintained over time (i.e., Transformational Information Properties).

Once the TIPs are defined, the archive should capture the values of the Information Properties and preserve them so that if/when a Transformation is carried out then the Information Properties can be compared with the preserved values.

To make this a little more concrete, consider the following examples.

A TIP for a document may be the colour of certain text. The meaning of the red colour would be provided by the Semantic Representation Information. Of course, the Structure Representation Information will be different between the original object and the Transformed object. Transforming from a Word file to a PDF the test would be that the text which is red in Word is also red in the PDF – and it would make sense if the rest of the text is not.

Considering another Transformation, this time of a FITS file converted to a CDF[45] file. Again, the bit sequences will have been changed extensively. In such a case it could be asked how a curator could have satisfied himself or herself that the object as transformed had not lost required information content and therefore was still being adequately preserved. This is the way in which the curator would see the new object has continued to maintain authenticity.

The FITS file might contain an image; the CDF file should contain a similar image. However, just comparing the two images rendered on screens would be inadequate for scientific purposes. Instead, the curator would need to be satisfied, for example, that the data values of the pixel elements were identical in the two images at corresponding points; that the co-ordinates associated with each pixel in the two images were identical, for example the same latitude and longitude; that the units associated with the numerical values were the same in both images.

Science data is largely numerical or documentary. In a transformation the way in which the numbers are encoded may change, for example from an IEEE real to a scaled integer. In such a case a number in the old and the new formats should be the same to within rounding errors or predefined accuracy. Alternatively, co-ordinate system transformations may also require changes to the numerical values, which however should be reversible. Thus, the validity of the transformation in preserving these significant data values is testable.

The information about the testing of the Transformational Information Properties, and the values, both original and new, should be added to the Provenance, to provide evidence with which authenticity can be judged.

5.4 Types of evidence to support claims of Authenticity

There are many types of evidence which may be presented to support claims of Authenticity including:

- Technical provenance including:
    - Evidence that the bits have not been changed such as:
        - Fixity Information in a digital signature or seal e.g., hashes with associated signatures and certificates associated with a document, for example via a container
            - Evidence the hashes have not been changed such as

[45] See the CDF Frequently Asked Questions https://cdf.gsfc.nasa.gov/html/faq.html

                - Additional qualified electronic timestamps and/or signatures/seals
                - Blockchain related e.g., Guardtime[46]
                - Hashes of hashes (hash-trees) e.g., ACE[47]
            - A verification report from a qualified trust service provider that the electronic signature/seal and associated document are/were authentic
        - Bit preservation error correction activities and logs
    - If there are alterations to the bits then one needs details of the actions which altered the bits with supporting evidence as to why the new object should be regarded as a replacement which should be considered authentic – as discussed in section 5.3.1
- Non-technical provenance such as reputation of individuals and of systems and organisations. These may include:
    - Evidence of conformance to recognised standards for information security such as ISO 27001
    - Evidence of conformance to recognised standards for trusted digital repositories such as ISO 16363
    - Evidence of conformance to legal standards for e-signatures and services such as Council Regulation (EU) No 910/2014
    - Membership of relevant professional bodies by the organisation and/or its staff.

5.5 Limitations of this chapter

It would be very convenient if Authenticity could just be a “yes this is Authentic” or “no this is not Authentic”, however this is not possible with digital objects which have been copied and re-copied. One can be very sure, indeed almost certain, that a digital object is what it is claimed to be, at least in the short to medium term. However, the longer the time the greater that probability that a Transformation will be needed. In that case checking the Transformational Informational Properties (TIPs) must be relied upon. The limitation on what is described in this chapter is the reliance on the number of and the definitions of those TIPs, which in turn depends upon the archive staff, perhaps with the help of the Producers. It is not possible, at the time of writing, to provide a way to determine whether there are enough, adequate, TIPs, nor to provide an exhaustive description of how to create TIPs.

[46] https://guardtime.com/

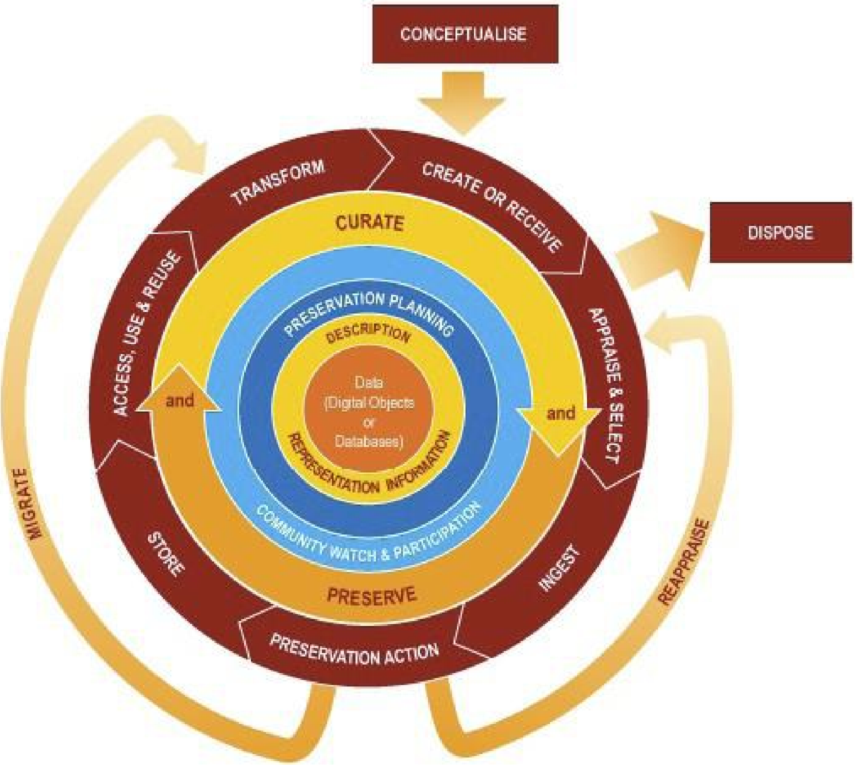
[47] https://wiki.umiacs.umd.edu/adapt/images/5/5b/DigCCurr2009_060909.pdf also see https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main

6 Lifecycles

It is possible to think about digital preservation in isolation i.e. I have been given this digitally encoded information and I must preserve it, but this leave open the question about where the information needed for preservation comes from. This chapter introduces various lifecycle models and some common features which are then taken up in the following chapter.

6.1 So many Lifecycle models

There are a large number of models for the "data lifecycle". The Data Life Cycle Models and Concepts document[48] collected together many data lifecycle models, and more recently Revisiting the Data Lifecycle with Big Data Curation[49]. Some examples are included below, ranging from very rich to the very rudimentary.



[48] CEOS Data Life Cycle Models and Concepts, CEOS.WGISS.DSIG.TN01 Issue 1.2 April 2012, https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS_Data-Lifecycle-Models-And-Concepts.pdf

[49] Pouchard, L., Revisiting the Data Lifecycle with Big Data Curation, IJDC, 2015, Vol. 10, Iss. 2, 176-192, http://www.ijdc.net/article/view/10.2.176

Figure 26 DCC Lifecycle Model http://www.dcc.ac.uk/resources/curation-lifecycle-model

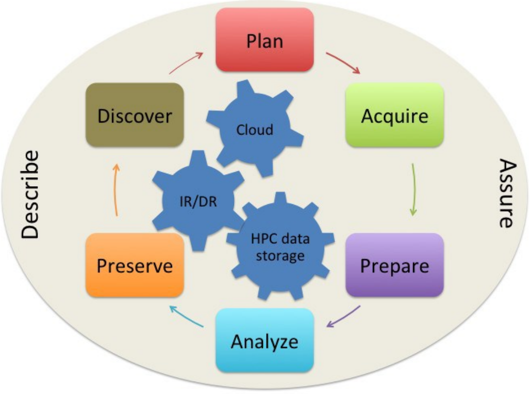


Figure 27 Big Data Life Cycle Model https://core.ac.uk/download/pdf/162675829.pdf

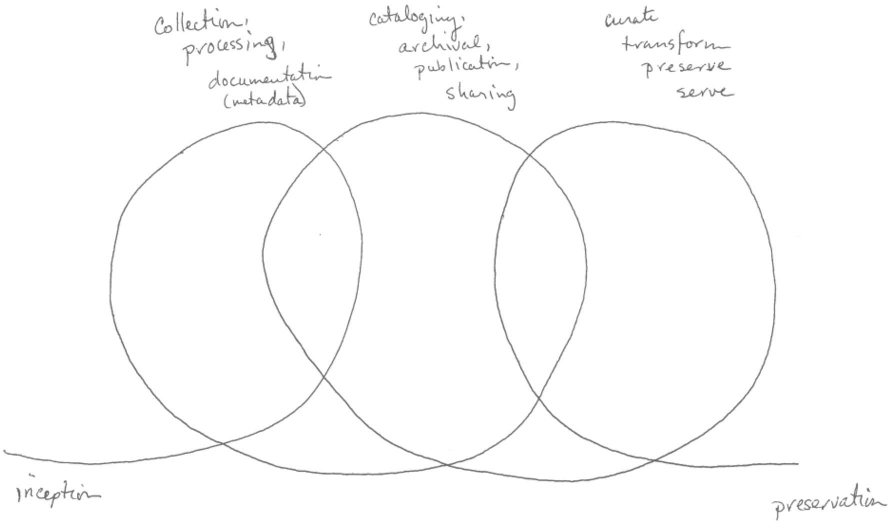


Figure 28 ELLYN MONTGOMERY, USGS, DATA LIFECYCLE DIAGRAM

All are some variations of the following, not necessarily sequential, steps:

Planning

- Acquiring
- Processing
- Analysing
- Preserving
- Discovering
- Accessing
- Re-using/re-processing
- Combining

This list includes all the FAIR principles,

Of primary importance in all these steps is what is generally called “metadata”. However, it is important to use a more detailed taxonomy, including those terms defined in the OAIS

Reference Model, in order to ensure that all the relevant types of metadata, with enough of each type, is collected along the way.

The document Information Preparation to Ensure Long Term Use[50] (IPELTU), which is in the process of being standardized by CCSDS and ISO, provides checklists for every stage of data production and use to ensure that the appropriate types of metadata are collected.

6.2 Limitations of this chapter

Rather than describing various models individually and in depth, it seemed more sensible to leave it to the reader to look at the documents for which the URLs are provided.

[50] CCSDS Publications see https://public.ccsds.org/Publications/MagentaBooks.aspx

7 Active Data Management Plans, collecting Information needed for Re-Use and Preservation – what to capture and when

If no thought is given to what is needed for its preservation and long-term re-use until digitally encoded information is handed over to a repository it is likely that a lot of the required additional information (metadata) will have been lost. This can happen because the documents and plans were regarded as unimportant and deleted or left to rot, either physically or digitally. As a result, the information being preserved will not be as useful and valuable as it might have been. Of course, one cannot collect every single scrap of information along the way, but this chapter will at least help people pause for thought and ask whether this or that piece of additional information might be worth keeping.

This chapter addresses the question of collecting the information needed for the preservation and re-use of target information. This might be achieved by accident[51]; however, such an approach cannot be relied on, and is not optimal. A planned approach may be described in two parts.

The first part may be called Data Management Plans (DMP), or as is preferred here, Active DMPs. This involves a rather high-level approach, largely driven by funders[52], which provide minimal requirements on DMPs. In order to provide a fuller analysis, the first section looks in more detail at the requirements for DMPs, not constrained by the current requirements of funders, to enable a more thought-through and thorough analysis.

The second part provides a more detailed identification of what should be collected, and when, which can be applied to the most general, complex, type of information creation.

7.1 (Active) Data Management Plans

Ensuring that information can be preserved and re-used does not happen by accident. There must be planning. Such plans are often called Data management Plans. Because such plans must evolve over time they are also referred to as Active Data Management Plans (ADMP).

[51] For example an important missing Ionosonde paper document was found propping up a desk – see Conway, E, CCLRC Ionosonde Case Study, 2007, http://www.alliancepermanentaccess.org/wp-content/uploads/temp/ionosonde-case-study.pdf

[52] See examples at DCC website https://www.dcc.ac.uk/resources/data-management-plans

These are increasingly mentioned and are, in many cases, required by funders. However, these tend to be static documents which do not evolve, cannot be tested or monitored and cannot be enforced. The ADMP process must:

- must be customized
- must be interoperable with other systems
- can evolve
- can be monitored
- can be enforced
- can be tested/ verified
- can be automated

The evolution of DMPs, as a project, which results in the creation and/or collection of information, progresses, could be along the lines of:

Rough ideas → Increasingly detailed plans → Detailed by evolving plans

For this high-level approach, based on what is required by funders, one can think in terms of:

a) Data
b) Metadata – used here as a collective term – finer level terminology is provided by OAIS.
c) Rights
d) Access/sharing
e) Archiving

Some ideas of the challenges and things that are likely to help are shown in the following table.

	Challenges	Contributions
Data	Can the data created be dealt with?	- Volumes of data are normally predictable, at least approximately.
Metadata	Have we got enough of the right type?	- OAIS information model: Representation Information, Authenticity evidence supported by preservation services. - Registries of various kinds - Rule based checking
Rights	Too restricted or not restricted enough?	- Various constraints including, commercial, financial, legal, and those specified by funders
Access/sharing	Are we using the “right” identifiers? Have we got the right discovery mechanisms? Are sharing mechanisms adequate?	- Some identifiers may be specified by project constraints, but additional identifiers may be assigned. - Can everyone who should be able to access and use the information, actually do so.

Archive	Where should we put the data and supporting information? What will the cost be?	- Audit and certification systems - Lists of repositories
(Adding) value	What is the probable/possible value, and to whom?	- Some sources of value may be clear from the information creators, other sources may be speculative or serendipitous.

Tools are needed to help the data creators develop and evolve the ADMP while at the same time help the funders, and related stakeholders such as the repositories, check that the plans are adequate. The text below provides more details based on experience with a number of types of data creation/collection including:

- manufacturing
- large national facilities
- astronomical satellites
- small group projects in cultural heritage, performing arts and various scientific disciplines
- individual researchers
- new data created by analysis of existing data
- cross-platform software systems

The following sub-sections expand on potential sources of information.

7.1.1 Monitoring the ADMP

An important part of the ADMP is that it can be monitored – so that any deficiencies in the ADMP can be detected and corrected by the data creators – who are probably the only people who know the required information.

The checking can be done in a number of ways, including those listed in the following sub-sections.

- **Checklists**

These can be done in an automated way – does some required thing, e.g., a document, exist. The quality of such a document would be difficult to judge automatically but some basic checks such as format or components of a complex object can probably be checked.

To judge the quality of the components of the ADMP is likely to require human judgment, and we might imagine that while some judgments require a significant amount of effort from a domain expert, nevertheless there may be some aspects which may be judged quickly by a non-expert.

- **Expert evaluators**

Disciplinary experts can perform more detailed checks. However, the APARSEN study[53] suggests that this will not be a normal occurrence.

- **Non expert evaluators**

A human could reasonably be expected to make some quality judgments by inspecting an image or a graph if there are some simple things that can be checked e.g., time stamps should increase monotonically, physical measurements are normally smoothly varying with respect to position or time. The data creator would need to provide the appropriate information.

  - **Simple display using data specific software**

The software may be specific to the data, but will show specific aspects to be inspected, although analysis software may be quite complex and require detailed documentation.

[53] http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2014/06/APARSEN-REP-D33_1A-01-1_1_incURN.pdf

- **Simple display using generic software**

Perhaps more useful for multidisciplinary use to ensure that the data is described in a way that allows access by generic software. What is shown in the diagram is the use of some types of Representation Information to describe the data. Examples include EAST and DRB, which can describe a wide variety of data; the data can then be parsed and used for other purposes. DRB could be likened as an annotated schema for a data set which can be used by something like XSLT to transform the data to something else.

- **Discovery**

The discoverability of the data is important – although this may only be possible after the data is published – for example when handed over to the repository. The check on discoverability may be broken down into:

- **Check Identifier**
  - Given an identifier the check is simply that something is there
  - Check searchability: the data creator should ensure that the data is findable.
  - Minimum check: the data creator provides the information about the search engine and search terms, and these should find the data
  - Multidisciplinary check: combine with data from other sources

7.1.2 Information from planning stages of the project

The planning stage is the initial phase. At this point most of the details of the data have not yet been decided. There are various aspects which will be readily available.

- **Current research information system (CRIS)**

The basic information about the project is normally held in a registry. A common exchange format is CERIF[54] provided by EuroCRIS[55] CERIF.

- **Planned rates/volumes**

Data rate and total volume are both important in determining problems to be addressed by the immediate data management plans. They are often linked by the length of time involved. The length of time for data collection may be well defined by the funding or else may depend on the success of the initial data collection.

Alternatively, the volume of data may be very well defined, for example a survey requiring a specific number of samples.

- **Levels of data**

The data collected or created is often termed raw or Level 0. This data may need to be processed to remove instrument signatures – perhaps requiring calibration data – producing what is often referred to as Level 1[56] data. Further processing may be needed, producing Level 2, Level 3 etc, before the data is usable for its primary purpose – and indeed may be essential before others can re-use the data. In many instances all the levels of data may be of interest, depending on the expertise of the next user.

At the planning stage it is likely that some details will be known about each of these levels, although the processing steps and calibration data may not be available.

- **Standards**

[54] Common European Research Information Format (CERIF) https://www.eurocris.org/cerif/main-features-cerif

[55] European CRIS https://eurocris.org/

[56] For example see NASA Data Processing Levels for Earth Observation systems at https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-information-policy/data-levels

The standards, whether general international, disciplinary, project specific, can be defined. For example, the project specific ones may be determined by an international collaboration, perhaps with special semantic conventions

- **Formats and semantics**

The basic formats will probably be reasonably well defined, as will be the basic semantics. There are many sources of recommendations[57], [58].

7.1.3 Creation

In this phase the data is collected/created. Experience suggests that in some cases the original plans will be followed perfectly; in other cases, there will be many details to be specified, including details of formats, semantics, error flags, modes of operation, ideas of potential interactions with other data, software, evolving standards, all possibly through several iterations.

- **Rates**

Rates of collection – or at least the maximum rate of collection must be determined

- **Volumes**

The volume of data should be well understood

- **Levels**

As described under Planning, there may be data processed through several levels, all of which may be important. In the creation phase the details of the software, processes, and additional information such as calibration data.

- **Production process between levels**

Experience suggests that processing between levels may be fixed from the start or, more likely, it evolves over time as the instruments and data are better understood. In the latter case there will be a number of versions. These versions may be important, with different versions associated with different datasets. Alternatively, all the data may be re-processed either as the final stage of the project or else on demand.

- **Process description**

The processing at each level may be simple or may be complex. In the latter case it consists of a number of smaller steps with decision points between each step or be performed in a monolithic piece of software. Besides software there may be other input, for example calibration data or human input.

The different components, software, or data may change independently.

- **Processing software**

Software can involve a mixture of languages, for example C, C++, C#, Java, Fortran, with some open source and others not, and software written in one language using libraries written in another. The software will run on one or more operating systems, perhaps distributed over many locations.

- **Provenance**

The provenance tells us when and where the data was processed, and how. In some cases, it is important to be able to validate the research by re-performing the data capture/creation.

[57] UK Data Services File Format: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/file-formats/

[58] US Library of Congress, Recommended Formats Statement, see https://www.loc.gov/preservation/resources/rfs/

For observational data it may not be possible to re-capture the information. However, it may be important to be able to re-run the processing.

Additional information is required and can be encoded in a variety of ways, for example OPM. Each of these encodings needs its own Representation Information.

- **Who did it**

Some kind of unique identification of the person or groups of people who created the data – or each level of the data.

- **When**

The time of data collection, processing events etc

Representation Information

7.1.4 Primary use

The data is normally collected for some initial analysis. This involves a number of other aspects which may be important for the data creator(s) and therefore important for the initial portion of the lifecycle of the data and the management of the data.

- **Analysis software**

This can be described in a similar way to the processing software discussed above.

- **Underlying concepts**

The concepts which are needed to understand the data and its analysis are difficult to capture explicitly but references to papers, terminology etc. can help. These are important for the definition of the Designated Community.

- **Related data sets**

To analyse the data may require other datasets. These datasets will have their own Representation Information and Provenance.

7.1.5 Handover (to a Repository)

The data will be put in some (perhaps many) repository(ies). They should be involved in decisions about managing, preserving, and re-using the data.

7.1.6 Re-use

An important aim of a data management plan is to help to ensure that the data can be re-used as easily as possible. Since no-one can foresee the future, it is not possible for the data management plan to be specific. Nevertheless, it would be useful to have at least some ideas from the people who know the data best. These ideas from the data creator(s) will also help with ideas about how best to publish the data. These ideas could include:

- Access restrictions
- Potential related uses
- Potential related data sets
- Compatible analysis software

  This is related to the virtualization, if any, that is available.
- Reproducibility

  An increasingly important issue is the reproducibility of the research. The information noted earlier will help ensure the reproducibility of the research to the greatest possible extent.

7.1.7 Preservation

Preservation is judged on the continued usability of the data. The discussion above covers most of the aspects needed, as described next.

- **Which repository**

A decision must be made before hand-over. There may be a list of approved repositories specified by the funders.

The second half of this chapter describes how plans to collect additional information can be applied to the most complex type of projects, in a systematic way.

7.2 Information Preparation to Enable Long Term Usability (z)

IPELTU is focussed on the Additional Information that needs to be captured and/or generated and retained in order to ensure that the information created by the project, either as part of its main objectives or as a by-product of achieving those objectives, can be exploited over the short, medium and long term. It is expected that, by ensuring this Additional Information is collected as fully as possible, projects can significantly improve their information legacy to the benefit of the wider community.

IPELTU uses a very general approach to describing projects, in terms of what are termed Collection Groups, namely “Initiating”, “Planning”, “Executing” and “Closing” in each Activity Area, each requiring Additional Information, which is the information that should accompany Data to ensure that it can be preserved and exploited. This will include Representation Information and Preservation Description Information (PDI), as defined by OAIS.

A project

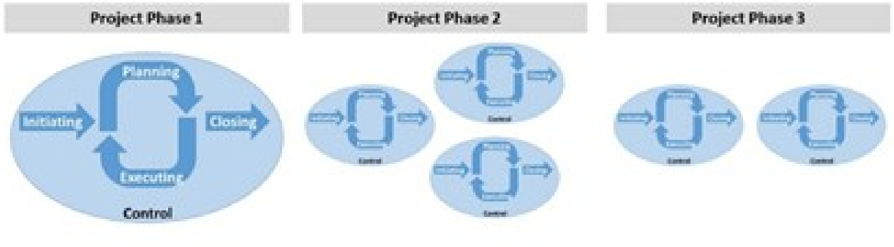


Figure 29 Phases and cycles in a project which collects/creates information to be preserved/curated

Table 2 provides examples for the various stages, from IPELTU. The IPELTU document provides further details and checklists for a number of types of projects.

7.2.1 Initiating Collection Group

The Initiating Collection Group consists of processes performed to justify the data collection and to define a new project, or new phase of an existing project, by obtaining authorization to start the project or phase.

This could include proposing the project/phase, perhaps responding to solicitations and funding information available. It would be reasonable to expect the following types of information to be created:

- the aims of the project to be clear enough to justify the data collection and its resources;

- the way in which data would be collected and the kind of data to be collected would be known in general terms;
- the initial exploitation of the data would be outlined.

These are likely to be important pieces of Additional Information that should be preserved as documentation of the project. The participants in this group of processes will almost certainly include sponsors and proposers and may also include data managers and archivists. Examples of documents to begin managing during project initiation include the list of project participants and organizations represented, the criteria for data collection, privacy and data protection, the criteria for repositories where the project data and documentation will be preserved, agreements among participants regarding authorship ownership of intellectual property produced by the project as well as relevant policies of participating organizations regarding such rights.

7.2.2 Planning Collection Group

The Planning Collection Group consists of those processes performed to establish the total scope of the effort, define, and refine the objectives, and develop the course of action required to attain those objectives.

In the Planning Collection Group, the preparations are made to collect or create data. This could include:

- the design and assembly of the components of the information system;
- the development or update of hardware and/or software systems;
- the development of the associated procedures for data collection, privacy, and protection;
- the establishment of a data dictionary.

These are likely to be important pieces of Additional Information that should be preserved as documentation of the project. Examples of documents to be managed during project planning include the project mission statement, the project management plan, the communication plan, the risk management plan, assignments for roles and responsibilities of team members, the list of project deliverables, the list of candidate repositories and how they meet the established criteria for managing data and documents produced by the project.

7.2.3 Executing Collection Group

The Executing Collection Group consists of those processes performed to complete the work defined in the project/phase plan to satisfy the specifications.

Activities are carried out which:

- create or collect the data;
- process and analyse data.

These processes will produce data that needs to be preserved for the long term either as a product or by-product. Examples of documents to be managed during project execution include signed contracts and approvals received from stakeholders or other authorities, data access policies and processes such as processing algorithms, validation and qualification plans, qualification matrixes, testing results, and project logs.

7.2.4 Closing Collection Group

The Closing Collection Group consists of those processes performed to conclude all activities across all Collection Groups to formally complete the project phase, or the entire project.

The data which may be part of the legacy of the project, and which can be exploited in various ways include:

- publication of research findings;
- generation of income;
- exchange of social information;
- predictions;
- scientific and social advancements.

There may also be ideas for exploitation in future.

The Closing Collection Group is performed by the project/phase team to use/re-use and exploit the information and, if appropriate, prepare it for handing over for long-term preservation, re-use, and exploitation. Examples of documents to be managed during project closing include signed acceptances, procurement documents, associated data, and related publication.

7.2.5 Control Collection Group

The Control Collection Group consists of those processes performed to ensure the project is on track or to identify areas which need attention. This process group provides information needed to manage the other process groups. The information collected during the controlling processes is part of the legacy of the project and therefore may need long term preservation.

This could include:

- programmatic changes;
- configuration management materials;
- changes in development or execution schedules;
- program or design review materials;
- changes in scope;
- test results.

Examples of documents to be managed during project monitoring and controlling include Configuration Change Requests and other documents describing proposed changes, and documented decisions of the Change Management Board or other decision bodies, test procedures and logs.	**Collection Group**			
Additional Information Areas	**Initiating**	**Planning**	**Executing**	**Closing**
Data Object	• Estimate of volume of data to be produced • Ideas of the potential value of the data	• Update Additional Information from Initiating based on more detailed plans • Identify types of data (raw, processed, etc.) which should be preserved • Identify types of data e.g., images, tables – and any generic interfaces • Quality constraints • Planned rate of data production • Expand and add detail	• Update Additional Information from Planning based on what really happens	• Finalise Additional Information from Executing • Inventory of data produced which should be preserved • Volume that would require preservation • Collect quality checks which may be performed on the data by non-experts • Define Information Properties which may be useful • Checks for (and logs of) any missing data

Representation Information	• Standards planned to be used • Information Model	• Update Additional Information from Initiating based on more detailed plans • Review applicable standards • Refine Information Model • Choice of data format • Identify Hardware and Software Dependencies • Relationships between data items	• Update Additional Information from Planning based on what really happens • Collect Semantics of the data elements e.g., data dictionaries and other semantics • Collect Format definitions and formal descriptions • Create Other Data Documentation • Calibration and system test tools and system test data that will be delivered	• Finalise Additional Information from Executing • Finalise Representation Information Networks to reasonable level • Identify other software which may be used on the data • Create suggestions for the Designated Community and Representation Information needed
Reference Information	• Identify standards which will be used to identify and reference the data and metadata	• Update Additional Information from Initiating based on more detailed plans • Identify which unique identifiers should be used (e.g., DOI or other)	• Update Additional Information from Planning based on what really happens • Rules, methods, tools for referencing data • Generate references to data as it is being created/captured	• Finalise Additional Information from Executing • Identify what may be used in future to identify the Information • Checks for (and logs of) missing references and logs of any

Provenance Information	• Record of origins of the project e.g., in a Current Research Information System (CRI)	• Update Additional Information from Initiating based on more detailed plans • Define Processing workflow, Processing inputs and Processing parameters • Define System Testing required • Documents from system development milestones	• Update Additional Information from Planning based on what really happens • Documentation about the hardware and software used to create the data, including a history of the changes in these over time • Update Documentation of Processing workflow, Processing inputs and Processing parameters • Record who was responsible for each stage of processing • Record when each stage was performed • Record of any special hardware needed • Record Calibration • Processing logs • Record checking of Fixity	• Finalise Additional Information from Executing • Finalise Provenance handover
Context Information	• Outline of background concepts needed to understand the project	• Update Additional Information from Initiating based on more detailed plans	• Update Additional Information from Planning based on what really happens • Collect publications related to the data or the processing system • Potential Value of the data and likely business case for sustainability	• Finalise Additional Information from Executing • Identify related data which may in the future be combined with this data

Fixity Information		• Fixity mechanism (e.g., CRC or digest) of data which may be preserved	• Update Additional Information from Planning based on what really happens • Identify any special validation procedures that should be carried out.	• Finalise Additional Information from Executing • Identify how do we verify that all files are intact
Access Rights Information		• What are the restrictions on access in the long term? • Clear identification of Intellectual Property Rights • Owners of the data – who can authorize hand-over	• Update Additional Information from Planning based on what really happens	• Finalise Additional Information from Executing • Licenses involved • The owner, and the restrictions on access (licenses), and the intellectual property rights
Packaging Information				• Details of the way components are packaged together for delivery to a repository • Definition of mechanisms for transferring information to next element in the workflow or next in the chain of preservation (e.g., definitions of SIPs)
Descriptive Information			• Identification of methods for exploration/ quick look at the data	• Finalise Additional Information from Executing • Create browse/query data if needed

Issues Outside the Information Model	• Estimated Cost of the project	• The budget for archiving and its relationship to the. overall budget for the project • The schedule for major project milestones and deliveries to the archive. • Identification of archives which are likely to be able to host the data	• Update Additional Information from Planning based on what really happens	• Finalise Additional Information from Executing • Schedule of deliveries • Pointers to the components to be transferred to the next element in the workflow or next in the chain of preservation • Potential preservation aims for the information created • Potential risks to preservation and exploitation of the data • Define the mechanism for communication between project and archive. • Define suggested Transformational Information Properties • Publications, or references to publications, including scientific publications, related to the project.

Table 2 Examples of activities in the collection groups at various stages Representation Information

Representation Information comes in many forms. It is whatever helps in the understandability and use of the data object i.e., the bits. There are many types of Representation Information which can be usefully classified as Structure, Semantics and Other. The next sections describe each of these in turn.

7.2.5.1 Structure Representation Information

This is the Representation Information that imparts information about the arrangement of and the organization of the parts or elements of the Data Object.

NOTE: For example, Structure Representation Information maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

7.2.5.2 Semantic Representation Information

The Representation Information that imparts information about the arrangement of and the organization of the parts or elements of the Data Object.

NOTE: For example, Structure Representation Information maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

7.2.5.3 Other Representation Information

A type of Representation Information which cannot easily be classified as Structure Representation Information or Semantic Representation Information. It is a type of Information Object.

NOTE: For example, software, algorithms, encryption, written instructions, and many other things may be needed to understand the Content Data Object in ways exemplified by the Preservation Objectives, all of which therefore would be, by definition, Representation Information, yet would not obviously be either Structure Representation Information or Semantic Representation. Information defining how the Structure Representation Information and the Semantic Representation Information relate to each other, or software needed to process a database file would also be regarded as Other Representation Information.

7.2.6 Provenance Information

Provenance information is important if one is to be able to reproduce the object's creation and processing.

For a specific data object there are likely to be many "siblings" with similar histories, but the most recent activities will be unique to that object.

The Provenance information would initially be relatively simple until information from multiple sources are combined e.g. a FITS image for an area of the sky in one wavelength combined with similar images for other wavelengths. Handling such Provenance is still an area of active research see https://www.sciencedirect.com/topics/computer-science/data-provenance and . https://www.research.ed.ac.uk/en/publications/data-provenance

There are multiple methods in use to capture Provenance, often dependent on the domain. For example, the library community often use PREMIS, which can be associated with specific vocabularies. Open Provenance Model (OPM, https://openprovenance.org/) and PROV (https://www.w3.org/TR/prov-dm/) are available standards.

Many scientific formats contain elements of Provenance, for example FITS files https://www.loc.gov/preservation/digital/formats/fdd/fdd000317.shtml have COMMENT and HISTORY records which can be used to describe the data unit and its provenance, while can also contain customized Provenance. In such cases the way in which the Provenance is encoded may be specific to a particular project and specific code would be required to extract it, perhaps into one of the standard formats to allow efficient queries to be performed.

Where the Provenance, or part of it, is encoded within the Data Object, as with FITS or HDF, the Provenance Information should describe how to extract that (piece of) the Provenance and its Representation Information, such as its Semantic RepInfo.

The Provenance Information may be encoded as a separate Data Object, for example PREMIS, OPM etc., and therefore there must be associated Representation Information for that object, for example the definition of the version of PREMIS and specific vocabulary used.

7.2.7 Context Information

Context Information is the information that documents the relationships of a Data Object to its environment. This includes why the Data Object was created and how it relates to other Data Objects.

Context Information can be a simple text document which describes why the data was created, for example as part of a scientific project. Provenance is a type of Context. Other examples include:

- Calibration history
- Related data sets
- Mission
- Funding history

7.2.8 Reference Information

Reference Information is the information that is used as an identifier for a Data Object. It also includes identifiers that allow outside systems to refer unambiguously to a particular Data Object.

It identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Data Object. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Data Object. Examples of these systems include taxonomic systems, reference systems and registration systems. In the OAIS Reference Model most if not all of this information is replicated in Package Descriptions, which enable Consumers to access Information of interest.

For scientific data examples include:

- Object identifier, including one or more types of Persistent Identifiers
- Journal reference
- Mission, instrument, title, attribute set

7.2.9 Descriptive Information

Descriptive Information allows users to locate information of potential interest, analyse that information, and order desired information.

Descriptive Information contains the data that serves as the input to documents or applications called Access Aids. The Descriptive Information is generally derived from the Package Description, which is itself derived from the Content Information and PDI. The Descriptive Information can be viewed as an index to enable efficient access to the associated Information Package via associated Access Aids. Access Aids are documents or applications that can be used to locate, analyse, retrieve, or order information from the archive.

7.2.10 Packaging Information

This is the information that describes how the components of an Information Package are logically or physically bound together and how to identify and extract the components.

An information Package is quite a general object, and the Packaging Information describes how the various components are arranged and how they can be extracted.

OAIS does not provide any details for the SIP or DIP. On the other hand, OAIS provides a great deal of requirements for the Archival Information Package (AIP), and so the bulk of the following is about the AIP.

Only the XFDU packaging format maps directly to the OAIS Archival Information Package.

An OAIS AIP may be as simple as a collection of pointers to the various components (see section 2.4) required by OAIS, namely

- Data Object
- Representation Information
- Provenance Information
- Fixity Information
- Reference Information
- Context Information
- Access Rights Information

The packaging Information would in such as case be definition of the way in which the pointers may be identified

A ZIP or TAR or other general "box/holder" containing all the components in a flat structure could also be an AIP as long as it was accompanied by the instructions as to how to identify and extract the various required components, for example having such instructions in a simple text file.

On the other hand, despite the name, the E-ARK specification does not seem to be compatible with the OAIS AIP because it does not seem possible to assign a place for the various required components, for example the Representation Information.

7.2.11 Supplementary Information held by the archive

As long as the objects are being preserved by the archive the following should also be preserved:

- Definition of the Designated Communities;
- Preservation Objectives;
- Transformational Information Properties (checks of the values of these may be part of Provenance Information).

For as long as any specific AIP structure is in use:

- Packaging Information for the AIPs;
- Relationship between Editions and Versions of AIPs.

Other information may be useful while the various services are in use:

- Finding, Ordering, and Retrieval Aids;
- Packaging Information for SIPs, DIPs.

7.3 Limitations of this chapter

This chapter has summarised examined the requirements for Active Data Management Plans, supplemented by the details provided by IPELTU but there are many details and checklists which have not been included.