ESSENTIAL DIGITAL PRESERVATION PART 1:

WHY READ THIS BOOK? WHY IS PRESERVING DIGITALLY ENCODED INFORMATION IMPORTANT – BUT DIFFICULT?

DAVID GIARETTA

david@giaretta.org

www.giaretta.org and www.iso16363.org

This book is being published in 5 sections.

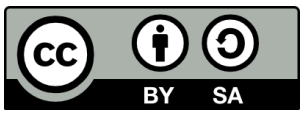
Part 1 : Why read this book? Why is preserving digitally encoded information important - but difficult.

Part 2 : Fundamental ideas about preserving digitally encoded information

Part 3: What to do and when to do it, to preserve digitally encoded information

Part 4: Adding Value and Exploiting Information

Part 5: Evaluating claims about preserving digitally encoded information



Contents

1 Introduction

Digital preservation is difficult. Many of the reasons this is so are well known, but others are less so. This book will explain these reasons and the ways in which these difficulties can be overcome.

Some of the details may be difficult for some readers. However, these details are important for those who wish to preserve digitally encoded information of all kinds, not just digital equivalents of paper.

Consider a motor mechanic working on your family's car. You would be worried if that mechanic did not know the details of how a car works including details of the engine and brakes. Similarly, you would not wish to fly in a plane for which the designer did not have detailed knowledge of the appropriate topics which require working diligently through, and mastering, very hard mathematical and engineering techniques. Even the pilot will need to know a fair amount about the fundamentals of flight in order to deal with emergencies. Lack of understanding such details is likely to be discovered fairly quickly through mechanical failures leading to deaths and injuries. The speed of discovery will allow corrections to be made to avoid further disasters.

Lack of understanding in the realm of digital preservation might not be discovered so quickly. It might be decades before failures are discovered, but the results could be widespread and disastrous in terms of social, organisational, or personal reputation or finances. The delay in discovery of the problem would mean that it is likely to be too late to remedy the problems because the missing information is likely to be irrecoverable. Also, the interrelationship between digitally encoded information may cause lead to the damage being extremely widespread.

In order to off-load the burden, in some, perhaps most, cases the responsibility is handed over to a software vendor of service provider. The question then is how can one select which system to use. The inevitable, but sad, truth is that those who wish to persuade you to use their software or services will not point out the limitations of their systems and may even mislead you in terms of their system's capabilities.

This book will equip the reader with the mental tools, so to speak, to ask the key questions which must be answered.

The advantage of this approach is that cookbooks, checklists, and lists of instructions become out of date rather quickly as technology changes, and very often they are limited in their scope.

The approach will be to start by describing some reasons why digital preservation is important and then working through, in a logical manner, how to preserve, using the OAIS

Reference Model[1], and building from there through the many challenges and opportunities into which these concepts lead us.

The phrase "digitally encoded information" is used rather than "digital data" or "digital objects" because the focus throughout is on preserving the "information" rather than simply "the bits". The shorter "digital information" is used at times in place of "digitally encoded information" for convenience in this book.

Some of the many reasons why digital preservation is important are described in chapter 0. Numerous examples of digital information are examined in considerable detail in chapter 4.

Some of the sections may need some persistence from the reader, especially for those who are steeped in the document preservation world, but that persistence will be rewarded. Digital documents are becoming more complex as links, apps, and other objects are embedded to make documents more useful, more updateable, and more user friendly. These enhanced documents will eventually arrive at libraries and archives for preservation, and someone needs to be able to understand what needs to be done.

Challenges that we must confront when preserving the digitally encoded information are described in later sections, leading directly into later sections, explaining why OAIS is the way it is and explaining what OAIS conformance actually requires. OAIS has recently been updated; these updates, which are used in this book, including the important concept of Authenticity, are described in later sections.

1.1 Why is this book different?

It will be noted that in this book the focus is on the details which will probably be unfamiliar to most people interested in digital preservation, rather than those things that are often talked about in the library and archive world, such as the applications needed to access documents, such as MS Word or Adobe PDF. Such digital documents would be considered to be successfully preserved if, after a number of years, they can be rendered, i.e. printed, or displayed, for viewing by humans. This is analogous to paper documents being readable if the paper is prevented from rotting away.

The view of digital preservation in terms is too limited because, as we will see in later sections, there are much more complex types of digitally encoded information, which will need to be preserved, but which are not normally simply printed out. Instead, they are used in other computer processes or combined with other information. Even Word and PDF are much more complex than one may think. Moreover, Word and PDF can contain or point to, and display, other rather complex objects like spreadsheets and databases.

Most of us are used to a world in which clicking on a file "automagically" causes an application to open which then displays that file's contents so that all we need to do is to look at and read it. We do not think about what "magic" happens behind the scenes to allow this to happen.

This "magic" actually relies on lots of clever software and configuration files which together make astonishingly good guesses about what we need to happen. These in turn rely on a vast army of software developers, some of whom do it as a paid job and others do it voluntarily.

[1] The OAIS standard may be downloaded for free from CCSDS at https://public.ccsds.org/Pubs/650x0m3.pdf and the same standard can be bought from ISO web site. Other details are available at http://www.iso16363.org/standards/iso-14721-oais/

What happens in future? One may decide that it is safe to assume that Microsoft and Adobe will take care of their customers and ensure that the "magic" continues. But can this really be relied on?

The approach in this book aims to be more general, as it needs to be because the digital objects that need to be preserved are not simple digital equivalents of paper

1.2 Background

This book collects and summarises knowledge gained from several decades of experience and research which can be summarized in the following way:

- Creating and managing scientific archives.
- Creating software for analysis and preservation.

Theoretical work on digital preservation of all types of digitally encoded information resulted in the OAIS reference model and its updates. The author of this book leads the working group which is responsible for writing and updating OAIS as well as ISO 16363, ISO 16919, and several other standards.

- A number of EU research projects:
  - The theoretical solutions were assessed in the CASPAR[2] project which used as examples many objects in each of scientific, cultural, and contemporary performing arts domains.
  - The requirements of users were sought in PARSE.Insight[3] which conducted probably the largest possible survey of researchers, publishers, and data managers across the world, across disciplines and backgrounds, to identify the most widely held concerns in digital preservation. There were enough responses to be sure of the generality of the conclusions because there was significant agreement between groups of disciplines and even groups across countries.
  - Specific challenges of linked data and ontologies were investigated in the PRELIDA[4] project.
  - The application of OAIS concepts to existing archives was conducted in the SCIDIP-ES[5] project.
  - Looking at the broader context of digital preservation, and in particular identifying how to justify, control and provide funding for preservation was undertaken in the APARSEN[6] project.

  Restrictions in the way that EU projects must be run, and the project structure required by the EU, meant that I could not do everything I would have liked in the

[2] CASPAR project – https://www.casparpreserves.eu/ or archived web content https://web.archive.org/web/20101209071123/http://www.casparpreserves.eu:80/ and https://web.archive.org/web/20220909120117/https://cordis.europa.eu/project/id/033572

[3] PARSE.Insight – https://cordis.europa.eu/project/id/223758 and archived web content https://web.archive.org/web/20100710085843/http://www.parse-insight.eu/

[4] PRELIDA project – https://cordis.europa.eu/project/id/600663 and archived web content https://web.archive.org/web/20141218044917/http://prelida.eu/

[5] SCIDIP-ES project – https://cordis.europa.eu/project/id/283401 and archived web content https://web.archive.org/web/20141023094942/http://www.scidip-es.eu/

[6] APARSEN project – https://cordis.europa.eu/project/id/269977 and http://www.alliancepermanentaccess.org/ and archived web content https://web.archive.org/web/20141218215959/http://www.alliancepermanentaccess.org/

projects mentioned here. The limitations include the distribution of people involved, the resource distribution, and the requirement for making claims of success. These meant that the projects achieved less than they could have, and, in some areas less than they claimed, especially in terms of future exploitation of results. Nevertheless, the projects did achieve very useful results as described here.

- Standards for the ISO Certification of archives (ISO 16363[7] and 16919) were created, and OAIS was updated, based on the lessons learned.

This book originated in the PTAB training course for auditors and repository managers[8] but covers a much wider range of topics. Much of this has been developed after the publication of the previous book, ***Advanced Digital Preservation***[9] although there is some overlap. There is a great deal of overlap with the earlier sections of my book ***Thinking Digital Preservation***[10], but there have been some updates to what was in that book. A full list of publications and presentations of the author is available at http://www.giaretta.org/david-giaretta/publications/.

I will apologise in advance to the purists who insist that one should use "data are" instead of "data is," because data is/are plural. I have used either "data is" or "data are," choosing whichever phrase seems more natural in the context.

1.3 Limitations of this chapter

The mental tools, as will be explained in detail in subsequent chapters, will not provide all the answers, but they should at least allow the reader to ask useful questions and evaluate answers.

[7] **ISO 16363** related information is available at http://www.iso16363.org

[8] See http://www.iso16363.org/courses/outline-of-modules-for-5-day-high-level-training-course-on-iso16363-for-auditors-and-managers-of-digital-repositories/

[9] Advanced Digital Preservation http://link.springer.com/book/10.1007%2F978-3-642-16809-3

[10] Thinking Digital Preservation (2022) see https://www.amazon.co.uk/dp/B0BK39T9ZPA

2 Why is Digital Preservation important?

Resources, whether money, people, software, or hardware, are needed to preserve digitally encoded information. These resources must be justified to those who provide those resources. This chapter provides rationales which may be appropriate for your work right now.

2.1 Which digital information should be preserved?

Information exchange is vital to human civilization. Verbal communication is ephemeral but can be repeated, and thereby retained, through human memory and further person-to-person communication, although there will be a risk of gradual change over the years. Many of these exchanges are of value, for some reason or another, to some people, who will wish to keep them for some period which may be days to years or more.

Drawings on walls persist for many tens of thousands of years. Documents written on vellum have lasted for more than 1000 years. Paper documents and drawings can be left on a shelf – as long as appropriate environmental controls are in place – for hundreds of years[11]. For the past few decades' information has been kept in a digital form.

Digitally encoded information, however, is fragile[12]. This will be discussed in more detail in chapter 4, but a simple example will suffice. Words on a piece of paper will decay slowly, unless the whole paper burns or is otherwise lost. The edges may crumble, and pieces of ink may be lost, yet the words will be readable. On the other hand, loss of a single bit may be catastrophic – the document may not be readable by the normal software[13]. Technology changes, both hardware and software, will change over even a few years.

Despite these, and the other difficulties which will be explored later in this document, it is important to preserve such information. The following sections provide some reasons why this is the case.

2.2 Data we are sure will be needed in the future

There are many items of data which the creators know will be needed in the future, whether in 5 years, 10 years, 50 years, or 500 years, or more. For example, the following clearly must be preserved.

- the records of how nuclear waste, which will remain radioactive for tens of thousands of years, has been disposed of – by processing and burying - must be kept, in case new and better methods of disposal are created;

[11] https://www.loc.gov/preservation/care/deterioratebrochure.html

[12] Wieringa, M., 2017. The Fragility of Digital Media Content: On Preservation and Loss: Sketching the Pilgrimage of Future Scholars to Recover Our Digital Vellum. Junctions: Graduate Journal of the Humanities, 2(2), pp.27–38. DOI: http://doi.org/10.33391/jgjh.33

[13] See Heydegger, Volker. (2009). Just One Bit in a Million: On the Effects of Data Corruption in Files. 5714. 315-326. 10.1007/978-3-642-04346-8_31.

- formula and details of processes for creating pharmaceuticals must not be lost because large future income streams depend upon them;
- the personal digital images and recordings created by individuals are valuable to each of those individuals, and to their families and descendants;
- financial information;
- contracts;
- registrations of births, deaths, and marriages;
- registration of land ownership;
- medical records
- evidence from criminal cases.

All these and more, certainly must be preserved for decades or centuries. Some of these objects will be physical, but here we are concerned with digital versions, either born-digital or digitized.

2.3 Legal requirements

We restrict this overview of legal requirements for digital preservation to sectors which are:

- high economic value and/or growth of the sector
- with a sophisticated IT infrastructure and
- sizeable international markets activities

Much of the following is taken from The Iron Mountain summary of European document retention guide (2013)[14] and the APARSEN Project[15]. We see those legal requirements vary from country to country and even between US states. Nevertheless, we see that much valuable information is legally required to be kept for many decades, and even longer.

The information identified in the following tables must be preserved for at least as long as the retention period, if only to avoid fines or other legal punishment for failing in its mandatory duties.

2.3.1 Tax and Accounting Sector

TAX and ACCOUNTING RECORDS	AUT	BEL	DEU	DNK	ESP	FIN	FRA
General obligation of taxpayers to provide (upon request of the tax inspector) all information that may be relevant to their tax position. , including all books, records, and other data carriers.	7y	7y	6-10y	10y	10y	6-10y	6y
A company is obliged to keep records of all delivery of goods or services, all intra-European Community acquisitions, all imports and exports, and all other information relevant for VAT purposes.	7y	7y	10y	5y	10y	6y	3y*
General obligation to keep at least the following records: (i) VAT invoices sent and received; (ii) documentation	7y	7y	10y	5y	10y	6y	3y*

[14] https://www.project-consult.de/files/Iron%20Mountain%20Guide%202013%20European%20Retention%20Periods.pdf

[15] See for example pp28-30 of APARSEN deliverable D11.6 http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2015/02/APARSEN-REP-D11_6-01-1_4.pdf

related to supplies and acquisitions within the EU; (iii) documentation relating to goods imported from, and exported to, outside the EU.							
Obligations to keep records and other data carriers relating to the ownership of real estate and rights relating thereto.	22y	-	-	10y	10-13y	13y	6y
Taxpayers are obliged to have available all information that deals with intra-group price setting, so that the Tax Authorities can check transfer and other conditions agreed upon in intra-group transactions	7y	7y	-	5y	8-18y	6-10y	3y*
Obligation to have administration showing the payment of dividends, and the obligation to issue dividend notes.	7y	7-10y	-	-	10y	6-10y	3y*
Keep and administration as per the requirements set out in the General Customs Act, including books, records, and other data carriers. General obligation of taxpayers to provide (upon request of the customs authority) all information which may be relevant to them, including making available all books, records, and other data carriers.	3y	-	-	-	3y	S	3y
Regional and municipal taxes	7y*	S	S	-	10y	S	1-3y*
LEGEND: Black text: minimum retention period. *: Recommended retention period. X: Retention prohibited. D: Duration of the contract or payroll. P: For the entire duration of the company and/or permanently. R: As long as required for the relevant purpose. S: Situation specific or too detailed for the scope of the summary and/or guide.							

Table 1 Examples of retention periods for tax & accounting records in Europe part 1

TAX and ACCOUNTING RECORDS	HUN	IRL	ITA	NLD	POL	ROU	SWE	UK
General obligation of taxpayers to provide (upon request of the tax inspector) all information that may be relevant to their tax position. , including all books, records, and other data carriers.	7y	6y	4-10y	7y	5y	5-10y	7y	1-6y
A company is obliged to keep records of all delivery of goods or services, all intra-European Community acquisitions, all imports and exports, and all other information relevant for VAT purposes.	7y	6y	10y	7y	5y	10y	7y	6y
General obligation to keep at least the following records: (i) VAT invoices sent and received; (ii) documentation related to supplies and acquisitions within the EU; (iii) documentation relating to goods imported from, and exported to, outside the EU.	7y	6y	10y	7y	5y	10y	7y	6y
Obligations to keep records and other data carriers relating to the ownership of real estate and rights relating thereto.	-	D+6y	10y	9y	-	10y	-	-
Taxpayers are obliged to have available all information that deals with intra-group price setting,	7y	6y	4-5y	7y	5y	5-10y	7y	-

so that the Tax Authorities can check transfer and other conditions agreed upon in intra-group transactions								
Obligation to have administration showing the payment of dividends, and the obligation to issue dividend notes.	-	S	10y	-	-	10y	-	-
Keep and administration as per the requirements set out in the General Customs Act, including books, records, and other data carriers. General obligation of taxpayers to provide (upon request of the customs authority) all information which may be relevant to them, including making available all books, records, and other data carriers.	7y	3y	4-5y	7y	5y	10y	5y	3y
Regional and municipal taxes	7y	-	4-5y	S	5y	10y	-	-
LEGEND: Black text: minimum retention period. BLUE text maximum retention period. *: Recommended retention period. X: Retention prohibited. D: Duration of the contract or payroll. P: For the entire duration of the company and/or permanently. R: As long as required for the relevant purpose. S: Situation specific or too detailed for the scope of the summary and/or guide.								

Table 2 Examples of retention periods for tax & accounting records in Europe part 2

2.3.2 Cultural and Creative Sector

Sub-sectors	Cultural Industries/Media & Entertainment: Film & video, TV & Radio, Video Games, Music, Books & Press (Publishing) Creative Industries: Design (fashion, graphic, interior, product designs) Heritage: Museums, Libraries, Archives & Archaeological sites Other core arts: visual and performing arts
Compliance	20-50Years for music, prototypes, and designs, +100 for long tail (e.g., film, cultural heritage) IPRs (copy rights and trademarks). Activities based on massive reproduction
Challenges	Transition to the digital content era Extract, combine and manage external and internal data Manage a complex cooperation / collaboration environment with new entrants and social media How to provide value-added services to potential customers (based on content) Maintain quality as a competitive advantage Create dynamic and interactive experiences based on existent content
Revenues	Cultural and creative sector turnover €541 billion 2003 (EU15) 3.5% of the GNP of the EU (2010) – cordis Broadcasting: €317billion 2014 to €405 billion in 2018 (+28%) Spending in ICT 12%

	Global broadcasting and cable TV industry expected to reach almost €332 billion in 2015 (MarketLine), 27% market growth in 5 years. TV advertising accounts for almost 48% of the overall market. Supports tourism: European arrivals reached 500 million. Tourism directly contributes, on average, 4.2% of GDP and 5.4% of employment (4.4% and 5.7% for EU members) in 2010.
Other	5.8 million of people worked in this industry 3.8 % of EU workforce (cordis) Creative industries are dominated by SMEs, with micro-SMEs and free-lancers representing 85% of all actors. SMEs co-exist with a few "global players," especially in publishing. The most important reason for publishers for DP: it will stimulate the advancement of science. Even most (96%) of small publishers DP is important. 84% of large and 55% of small publishers have a DP policy but 70% of them (large + small) do not have DP in place [16] Open to innovative ICT adoptions Shortages in IT budgets; do more with less

Table 3 Examples of retention periods for cultural and creative sector

2.3.3 Energy and Utilities

Sub-sectors	
Compliance	Permanent retention – Energy database and pesticides database 3-20Y Copies of waste management 30Y Documents containing audits on radioactivity and results measurement 10Y data regarding chemicals or environmentally dangerous substances 10Y Metering database
Challenges	Regulation for utilities / energy industry continues to growth to improve industry security and reduce risks Market transparency and exchange of information regulation Customer and regulators are looking for ways to reduce energy costs Outage Prevention, Readiness, and Response is a major priority IoT (Internet of Things) business value is yet to be seen Capture the opportunity that represents HEM (Home Energy Management) before other competitors Sustainability is major concern for cities and citizens Every record has both content and metadata for indexing, searching, and formal auditability. For one company, there are typically 1,000 users and 6 million documents of distinct types and sizes to be managed on a daily basis regarding nuclear operations Increasing need of interconnecting countries and grids increase energy security and efficiency
Revenues	According to the World Nuclear Association, 435 nuclear reactors exist in 30 countries and generate 14% of the globe's electricity

[16] PARSE.Insight D3.4 P5/83

	Globally, demand for electricity is set to continue to grow faster than for any other final form of energy. Demand expands by over 70% between 2010 and 2035, or 2.2% per year on average. In terms of electricity use, industry remains the largest end-use sector through 2035 The electricity sector's annual turnover of €420 billion represents more than 3% of European GDP (Electricity without Borders a plan to make the internal market work – BRUEGEL BLUEPRINT SERIES, 2013.)
Demand & Others	To maintain safety, security, and compliance at power plants, Management must have well-documented and highly visible information across the asset life cycle. Photovoltaic and wind energy capacity increasing

Table 4 Examples of retention periods for Energy and utilities

2.3.4 Healthcare

Sub-sectors	Hospitals, medical technology, and devices
Compliance	Highly regulated Patient lifetime Hospital safety records (i.e., incidents) 7-10 years X ray 30 years Ultrasound records (e.g., vascular, obstetric 20 years or 8 after death Post-mortem Registers 30 years[17]
Challenges	Dealing with cost pressures Cope with more regulations Providing services under staff shortages Maintaining safety and quality of service New Business Models Drive at least 50% of Healthcare IT Growth Compliance will cost more than expected by players of the sector
Revenues	Public expenditure on healthcare in the EU 14% in 2030 UK £20 billion is exactly the extra money that the NHS will need every year by 2020 to meet patient demand
Other	IT spending: in western Europe were above 6% for 2014 Shortages in IT budgets; do more with less

Table 5 Examples of retention periods for Healthcare sector

2.3.5 Manufacturing

Sub-sectors	Automotive, Aerospace, Discrete manufacture (no automotive), chemical and process manufacturing, Food
Compliance	Highly regulated +50 years for design Automotive: +15 years for vehicles sold Aerospace: +50 Y

[17] http://www.hse.ie/eng/services/list/3/hospitals/ulh/staff/resources/pppgs/rm/recret2013.pdf

	Discrete manufacturing: 15-50 Y Chemical and process manufacturing: safety 20-50 years Food: 1-30Y (safety)
Challenges	Create globally integrated value chains (+ integration and discovery) Enablement of connected operations Move beyond automation, transform labour force into knowledge worker Decoupling of manufacturing and management functions in geographically dispersed manufacturing companies Move beyond automation, transform labour force into knowledge worker Increasing need of big data and analytics
Revenues	€271 billion -2001 EU manufacturing 7.4 % growth (2010) EU chemicals 10.2 % growth rate (2010) EU chemicals trade €41.7 billion (2011) Food & drink: €1,117 billion (2012, +6.8% than 2010) . 287,000 companies in Europe. Aerospace: €26 billion worldwide (2012) Automotive: €780 billion, with value added of over €140 billion (2011)
Other	ICT spending €541 billion Gross value added of ICT in Europe almost 120 billion (2010) Food: 4.25 million people employed in EU (stable to 2012) Chemical: 1,19 million people employed in EU (2011) Automotive: 2 million people directly employed (2011) Shortages in IT budgets; do more with less

Table 6 Examples of retention periods for Manufacturing sector

2.3.6 Medical records - retention periods across Europe

Black text: minimum retention period. Blue text: maximum retention period.*: Recommended retention period. x: Retention prohibited. D: Duration of the contract or permit. P: For the entire duration of the company and/ or permanently. R: As long as required for the relevant purpose. S: Situation specific or too detailed for the scope of the summary and/or guide.

MEDICAL/ SAFETY RECORDS	AUT	BEL	DEU	DNK	ESP	FIN	FRA
Medical (occupational health and safety company doctor) files; medical documents in cases of a medical treatment contract	30y	15-40y	10y	10y	P+5y R	S	x
Floor plans and directions	-	-	-	-	R+5y R	R	5y

Work-related medical examinations related to hazardous substances	-	30-40y	40y	40y	P+5y R	-	X
List of employees who have worked under dangerous conditions or whose health has otherwise been under threat	-	-	-	-	P+5y R	-	P
Register of employees who work with 3rd and 4th category biological agents	-	10-30y	D D	10y	10-40y	40y	10y
Lists/register of employees who have been exposed to asbestos dust	40y	40y	40y	40y	40y	80y	P
Administration concerning measurements of radioactive substances	-	5y	-	-	20y	-	P
Records of radiation	-	30y	30y 100y	-	30y	5y R	P
Medical records of employees who have possibly been exposed to ionizing radiation	-	-	30y 100y	30y	30y	30y	P
Registration of work and rest periods (in appropriate format)	-	R	2y	1y R	D+4y R	2y R	D
Necessary data for emergency medical care, individual reintegration plans, individual treatment agreements, degree of incapacity for work, required workplace adaptations	-		R	10y R	D+4y R	R	D

Table 7 Examples of retention periods for medical records in Europe part 1

MEDICAL/ SAFETY RECORDS	HUN	IRL	ITA	NLD	POL	ROU	SWE	UK
Medical (occupational health and safety company doctor) files; medical documents in cases of a medical treatment contract	30y	D+7y 40y	P	15y	20-50y	D	-	S R
Floor plans and directions	R	P	10y	R	2y	50y	-	S
Work-related medical examinations related to hazardous substances	30y	40y P	S	40y	50y	40y	10y	40y
List of employees who have worked under dangerous conditions or whose health has otherwise been under threat	50y	S	D+ 10 y	40y	40y	40y	40y	5-40y
Register of employees who work with 3rd and 4th category biological agents	-	10-40y	D+10y	10y	10y	40y		40y
Lists/register of employees who have been exposed to asbestos dust	50y	40y- P	40y	40y	40y	40y	40y	40y
Administration concerning measurements of radioactive substances	5y	S	D+10y	5y	3-5y	10y	-	S
Records of radiation	5y	5y-P	10y	5y	30y	10y	10y-P	

Medical records of employees who have possibly been exposed to ionizing radiation	50y	S	30y	30y	30y	40y	30y	5-40y
Registration of work and rest periods (in appropriate format)	R	3-7y R	R	1y 2y	3y D	R	3y R	2-3y R
Necessary data for emergency medical care, individual reintegration plans, individual treatment agreements, degree of incapacity for work, required workplace adaptations	R	D+7y-P R	R	R	20-50y	R	R	D+6y

Table 8 Examples of retention periods for medical records in Europe part 2

See http://www.healthit.gov/sites/default/files/appa7-1.pdf and http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435263/

2.3.7 Medical records (USA examples)

State	Medical doctors	Hospitals
Alabama	As long as may be necessary to treat the patient and for medical legal purposes.	5 years
Arizona	**Adult patients:** 6 years after the last date of services from the provider. **Minor patients:** 6 years after the last date of services from the provider, or until patient reaches the age of 21 whichever is longer	**Adult patients:** 6 years after the last date of services from the provider. **Minor patients:** 6 years after the last date of services from the provider, or until patient reaches the age of 21 whichever is longer.
Massachusetts	**Adult patients:** 7 years from the date of the last patient encounter. **Minor patients:** 7 years from date of last patient encounter or until the patient reaches the age of 9, whichever is longer	30 years after the discharge or the final treatment of the patient
New York	**Adult patients:** 6 years. **Minor patients:** 6 years and until 1 year after the minor reaches the age of 18	**Adult patients:** 6 years from the date of discharge. **Minor patients:** 6 years from the date of discharge or 3 years after the patient reaches 18 years (i.e., until patient turns 21), whichever is longer. **Deceased patients:** At least 6 years after death

Table 9 Examples of retention periods for medical records in USA states

2.3.8 International Retention Periods for business records

- **Argentina** – Under the Commercial Code certain accounting records (financial statements, ledgers, and journals) required to be retained until closure of business plus **10 years**.

- **Belgium** – Under the law relative to accounting and the yearly accounts of enterprises, ac - counting records / tax documentation required to be retained for **10 years**.
- **China** – Under Rules for the Implementation of the Law on the Administration of Tax Collection, accounting records / tax documentation are required to be retained for **10 years**.
- **Czech Republic** – Under the Accounting Act, pension records and wage records supporting pension benefits are required to be retained for **20 years**.
- **Germany** – Under the Commercial Code, certain accounting records are required to be retained for **10 years**.
- **Japan** – The Commercial Code requires accounting and certain other business records to be retained for **10 years**.

2.3.9 Court proceedings and legal evidence

The courts themselves, to which breaches of any of the above requirements might be referred, are themselves creators and users of digitally encoded information. The proceedings and evidence presented in such court cases help to form the basis on which precedents and/or appeals may be based and are therefore of critical importance to individuals, companies, and governments.

Yet proceedings[18] in court have been marked as "Endangered," and the evidence in court[19] as "Critically Endangered." This concern extends to things to which courts refer. For example, between 1996 and 2013 US Supreme Court justices have cited materials found on the Internet 555 times ... but half of the links in all Supreme Court opinions no longer work[20].

Digital Forensics[21] involving computers and mobile devices is increasingly important in court cases. This has been described[22] as follows:

> *The majority of crimes currently have a digital component, such that Governments and the police are obliged by law to indefinitely hold digital evidence for a case's history. Until the presentation of the digital evidence in Court, the evidence must be collected, preserved, and properly distributed.*

It must be remembered that digital forensics requires more than just digital preservation since the original hardware and issues such as the contents of volatile memory come into play, however the results of such investigations must be preserved.

Some countries are taking real action, for example the Supreme Court of India has published its *DIGITAL PRESERVATION Standard Operating Procedure (SOP)*[23]. This

---

[18] In the 'Bit List' of Digitally Endangered Species, DPC, 2024 see https://www.dpconline.org/digipres/champion-digital-preservation/bit-list/endangered/bitlist-proceedings-in-court

[19] In the 'Bit List' of Digitally Endangered Species, DPC, 2024 see https://www.dpconline.org/digipres/champion-digital-preservation/bit-list/critically-endangered/bitlist-evidence-in-court

[20] DIGITAL PRESERVATION Standard Operating Procedure (SOP), 2021, E-Committee, Supreme Court of India, http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/

[21] Digital Evidence Preservation – Digital Forensics, see https://www.geeksforgeeks.org/digital-evidence-preservation-digital-forensics/

[22] Molina, Fernando & Rodriguez, Glen. (2017). The preservation of digital evidence and its admissibility in the court. International Journal of Electronic Security and Digital Forensics. 9. 1. 10.1504/IJESDF.2017.081749, https://www.researchgate.net/publication/312934498_The_preservation_of_digital_evidence_and_its_admissibility_in_the_court

[23] See https://districts.ecourts.gov.in/sites/default/files/sop%20compressed_0.pdf

document calls for ISO 16363 certification (see http://www.iso16363.org) of the repositories, which it shows as:

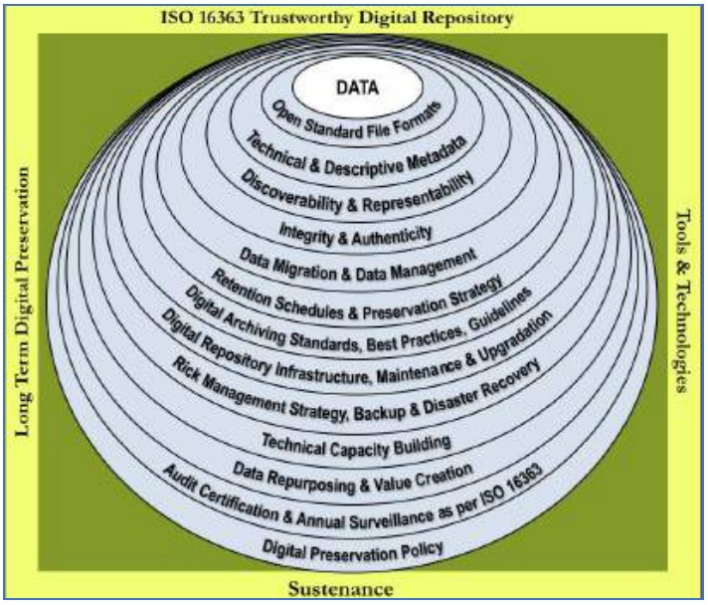


Figure 1 Layers of ISO 16363 Certified Trustworthy Digital Repository

India has 25 High Courts and approximately 672 District Courts; the SOP proposes an implementation shown in Figure 2.

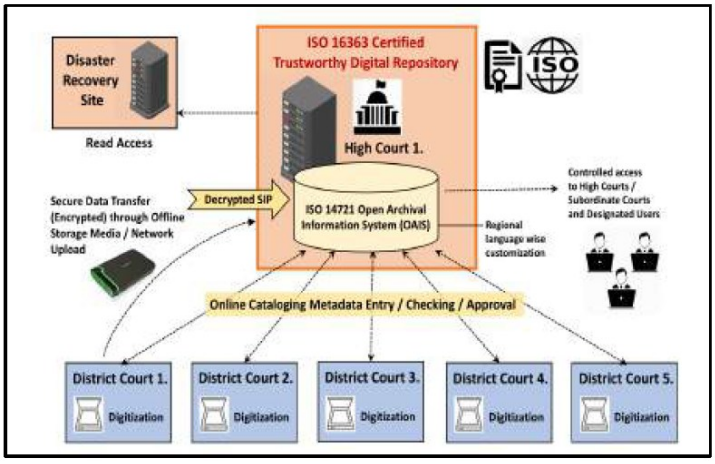


Figure 2 Implementation model for Indian Courts

The SOP stresses the critical importance of interoperability between the Judicial Digital Repositories (JDRs) at metadata, data, and systems levels.

2.4 Just in case information is needed later

By 2025, the amount of data generated each day is expected to reach 463 exabytes globally[24]. Some of this data will certainly be very valuable, if only to avoid fines, in the future. Of the rest, some will certainly **not** be valuable and for others, one will not be sure

[24] https://seedscientific.com/how-much-data-is-created-every-day/

one way or another, as is illustrated in the figure below[25].

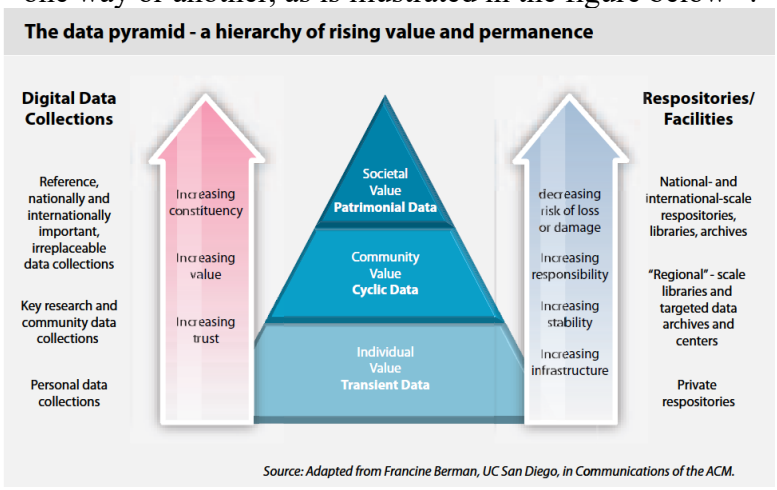


Figure 3 The data pyramid - a hierarchy of rising value and persistence

The problem is that it is difficult to know whether a specific piece of information will be valuable in future.

What should be done?

The data we are certain will be valuable, we should preserve, while the data which is certainly not valuable, we can disregard or delete. The data which are in the grey area – where one is not sure about its value - the sensible option would be to take steps to preserve it for some time, until its value becomes clear. Of course, one cannot leave things in limbo forever, but on the other hand there is less pressure to delete digital objects, which may be being stored on tape, than there would be to replace physical (paper) documents, which take up shelf space which is much more difficult to extend.

Whereas paper may be left on a shelf until a decision is made, digital objects must be actively preserved until a decision is made that it is not worth preserving.

2.5 Benefits of preserving data

Data takes many forms, including numbers and characters in tables, images, XML files to discipline specific file formats, as well as documents, publications, instrument designs, and many others. What needs to be preserved is the information encoded in that data.

The preservation of each type brings its own set of benefits for a variety of stakeholders.

A useful example is space-based research involving satellites carrying scientific instruments. Such missions cost vast amount of money, their primary purpose is to gather information, transmitted to the ground in digital form. The cost effectiveness of the mission, justifying the money spent can be difficult to judge in financial terms. However, a number of measures are commonly used.

In a small number of cases a single well-defined result is all that will be produced, but that is the exception.

The more usual case is that there will be some initial scientific results, often published by Principal Investigators (PI), the team that developed the instrumentation. Publication of any results from a PI mission is often seen as 100% effectiveness for the mission.

However, there are many instances in which the data is useful long after it is initially gathered. The most obvious case is long term monitoring, for example for Climate change

[25] Riding the Wave, 2010, https://www.fosteropenscience.eu/sites/default/files/pdf/831.pdf

studies, which may use data from many fresh sources, spanning the longest possible time. In other cases, different aspects, and different combinations, of the data are taken from that which was of interest to the PI, and other scientific papers are published well after those of the PI. Further scientific studies may use the same data in still different, ways at later stages. In the case of the International Ultraviolet Explorer (IUE), data is still used more than 40 years after it was initially obtained.

Measures of scientific effectiveness include:

- the number of papers written using the mission data - one could crudely argue that, say, doubling the number of scientific papers based on data from a particular mission would double the cost effectiveness of that mission.
- the number of references to papers drawing conclusions from the mission data - a single significant result on, for example the origin of the Universe, would be referred to by many subsequent papers.

Certainly, the former can be greatly influenced by expenditure on information exploitation. Information exploitation depends very heavily on making data accessible and usable to as large an audience as possible. However, this normally takes only a small portion of the project budget.

One important aspect of space missions, with their extreme demands on reliability and compactness is that of pump-priming for, and providing skilled staff to, industry. Even a mission that fails on launch will have some measure of success on these grounds. A very large portion of expenditure will go to industry in almost all cases. The same is true for "big science" such as High Energy Physics.

Data from many satellites, particularly that to do with Earth Observation, have commercial value. In many cases data gathered is unprocessed unless specifically requested at some point in the future.

Because research data is itself so diverse it is sensible to not try to restrict the types of information being considered, indeed it has been said[26] that one person's digital trash is another person's digital treasure.

Following this line of thought, some preserved information may attract commercial returns, for example attracting eyeballs for advertisements.

Other benefits include, pump priming for industry[27], national prestige and commercial exploitation of the information being preserved.

The following sections discusses the benefits of preservation from different viewpoints, with relevant examples of types of information, benefits, and motivations. Later sections discuss the exploitation of preserved information.

2.5.1 Stakeholders

It is worth looking at digital preservation from the point of view of the various stakeholders.

- Governments

[26] Homeland Security News Wire, 2013, see https://www.homelandsecuritynewswire.com/dr20130610-social-media-analytics-help-emergency-responders

[27] For example the Archiver project stated that its purpose was "ARCHIVER - Archiving and Preservation for Research Environments - will introduce significant improvements in the area of archiving and digital preservation services, supporting the IT requirements of European scientists and providing end-to-end archival and preservation services, cost-effective for data generated in the petabyte range with high, sustained ingest rates, in the context of scientific research projects." – see https://www.archiver-project.eu/about

    - Information of strategic value must be preserved, for example underground infrastructure such as pipes and communications conduits.
    - Information of national pride must be preserved, such as data which has been uniquely difficult to collect such as that from space missions.
    - Information on which policies are based should be preserved, for example historical land use and pollution, demographic trends, or results of previous policies.
- Multinational organisations
    - Multilateral agreements, such as precise locations of borders, must be preserved.
- Citizens
    - Information needed to hold the government to account must be preserved.
    - Information of long-term interest to the public should be preserved.
- Strategic Management and Funders
    - Information is costly to create/gather; it should be preserved to ensure it is not lost in order to ensure it does not have to be created/gathered again at similar cost.
    - Information is valuable so preserving its usefulness allows more value to be extracted.
    - Some information, e.g., measurements of climate, cannot be re-created so must be preserved for longitudinal studies.
    - Some information is too costly to re-create so must be preserved.
    - Some information must be preserved for legal reasons.
    - Digitally signed contracts for long term agreements must be preserved.
- Tactical Management
    - Information is fragile so steps must be taken to preserve it even over relatively short timescales in order to allow time to make a decision as to whether to preserve it over the longer term, or potentially forever.
- Investors, concerned with increasing value
    - Information is valuable as long as it is usable despite changes e.g., in technology.
    - Information can be combined together to become more valuable so preservation must enable this.

2.6 Motivations for an individual – the reader

Maslow[28] wrote:

[28] Maslow, A., 1943. A theory of human motivation. Psychological Review, 50(4), pp. 370-396.

"*It is quite true that man lives by bread alone—when there is no bread. But what happens to man's desires when there is plenty of bread and when his belly is chronically filled?*

At once, other (and "higher") needs emerge and these, rather than physiological hungers, dominate the organism. And when these in turn are satisfied, again new (and still "higher") needs emerge and so on. This is what we mean by saying that the basic human needs are organized into a hierarchy of relative prepotency."

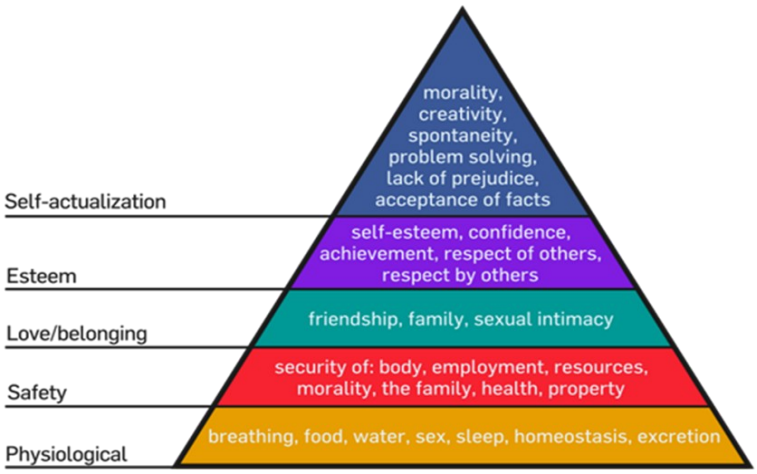


Figure 4 Maslow's Hierarchy of Needs

He then suggested a hierarchy of needs as illustrated in Figure 4.

At the base there are **Physiological Needs** which are the physical requirements for human survival. If these requirements are not met, the human body cannot function properly and will ultimately fail. Physiological needs are thought to be the most important; they should be met first.

For a digital preserver this may be interpreted as simply the need to be paid in order to buy food and shelter. In other words, you may be doing this simply as a job.

Safety Needs are next, which may come in many forms including job security i.e., if this digital information is lost then I lose my job. Or perhaps for protection against unilateral authority, for example, a Government can be held to account as long as this digital information is preserved. Alternatively, you might regard this as an insurance policy i.e., this digital object should be kept – just in case; this was discussed previously. The reason for preserving might be as a safety net against accidents/illness and their adverse impacts, for example this should be preserved because if "John Smith" retires then no-one will know what do. There is also fear of the unknown – we are not sure if this is worth preserving but just in case – similar to the previous discussion. Information preservation may also potentially satisfy a need for safety in a financial sense because it could be of commercial use e.g., it might be displayed to attract eyeballs to justify charges for advertising rates. Preservation may also be a safeguard such that preserving information will prevent legal problems, for example if there are legal requirements to keep that information safe, as discussed previously.

Above this is **Love and Belonging**, which can be satisfied, for example, by maintaining relationships with family, mentors, colleagues, confidants, by belonging to a group of fellow digital preservers, but there may be separate groups e.g., librarians vs archivists vs data curators. On a more individual level there is also the question of ***where do I come from?*** This could lead to a need to seek out, and preserve, family history, for your or your family's own use, using the ideas contained in this book.

There is then the need for **Esteem** which may be satisfied in a number of ways, for example there is a need for respect from others -perhaps related to digital preservation work that you done, which would be an incentive to do such work well, which this book will help with. You can demonstrate competence, proficiency in digital preservation by writing

research, publications, and reports, which are cited by others, and perhaps more significantly by ensuring that your archive is certified as ISO 16363 conformant.

At the top of the pyramid is the need for **Self-Actualisation**, which can involve seeking to reach full potential, wishing to keep everything that may affect that aim, in order to show "these are my accomplishments", and the evidence is your ability to preserve digitally encoded information.

Another level can be added, namely **Self-Transcendence**, perhaps the most important aspect motivating preservation, including preserving information for future generations, for posterity. This is also linked to the realisation of one's own mortality.

2.7 Horror Stories of information loss

US Supreme Court: Between 1996 and 2013 US Supreme Court justices have cited materials found on the Internet 555 times ... but half of the links in all Supreme Court opinions no longer work[29].

Global news: use of social media is becoming an important source of news ... but a study[30] showed that 11 per cent of the social media content had disappeared within a year and 27 per cent within 2 years (see Figure 5 from that study).

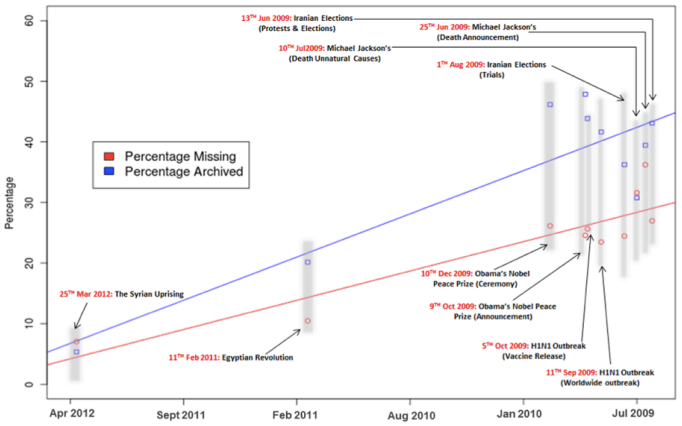


Figure 5 Percentage of content missing and archived for the events as a function of time

Even the **websites of major corporations** that should know better — including Adobe, IBM, and Intel — can be littered with broken links[31].

Scientific data: one of the foundations of the scientific method is the reproducibility of results ... but a survey[32] found the median lifespan of links in the scientific literature was 9.3 years, and just 62% were archived. Another survey[33] of 20-year-old studies shows that poor recordkeeping and inaccessible authors make 90 percent of raw data impossible to find.

It is difficult to find stories of significant losses of digitally encoded information because it is difficult to obtain admissions of loss/failure because they are so embarrassing to those responsible. However, the ***Atlas of Digital Damages***[34] has managed to collect a number of examples.

There are other stories which are less well founded but which appear in several places - often repeated but it is difficult to find original sources, and some are not actually as serious as it appears.

---

[29] http://blogs.law.harvard.edu/futureoftheinternet/2013/09/22/perma/

[30] http://arxiv.org/abs/1209.3026

[31] http://linktiger.com/broken-link-stats.php

[32] http://journalistsresource.org/studies/society/internet/website-linking-best-practices-media-online-publishers

[33] http://www.smithsonianmag.com/science-nature/the-vast-majority-of-raw-data-from-old-scientific-studies-may-now-be-missing-180948067/?no-ist

[34] See Barbara Sierman's Atlas of Digital Damages https://www.atlasofdigitaldamages.info/

1) The story, as reported many times[35], [36], is that a great deal of the US 1960 decennial census was lost - obsolete computer technology **BUT** this report is apocryphal as reported[37] later.
2) The story was that much of the data from the Viking mission to Mars was thought to have been lost[38] **BUT** actually it has been recovered[39].

It is always a good idea to check for updates on any horror story because attempts may be made at any time in the future to recover information, if there is sufficient interest. For example, the data from the SEASAT mission was considered lost. However, a resurgence of interest in the Earth Observation information from 1978 in order to analyse changes over the years by comparison of SEASAT images with e.g., more recent comparable ENVISAT ASAR[40] image data, led to recreation of its software and reprocessing of the data[41].

2.8 Disincentives – reasons against digital preservation

It is important to realise that although many of those reading this book will regard preserving our digital heritage as self-evident, nevertheless this is not universal opinion, as described in the next sub-sections.

2.8.1 Avoidance of legal, political, or criminal exposure

An organisation, or person, may be forced to reveal information they hold, for example by Freedom of Information[42] requests, or a legal discovery[43], [44] process in the courts. Such information may be embarrassing – or worse. Therefore, an argument may be made that it is best if information should not be written down or captured[45], [46], or be destroyed as soon as possible, for example using "triple deleting," in which an e-mail (or other digital object) is

[35] http://andromeda.rutgers.edu/~govdocs/stayingdigital.pdf

[36] https://www.atlasofdigitaldamages.info/stories/census-bureau-us/

[37] Adams, M.O. and Brown, T.E., Myths and Realities About the 1960 Census, 2000, http://www.archives.gov/publications/prologue/2000/winter/1960-census.html

[38] See https://www.atlasofdigitaldamages.info/stories/viking-lander-data/

[39] http://news.softpedia.com/news/Lost-Viking-Explorer-Data-Finally-Recovered-397746.shtml

[40] The Advanced Synthetic Aperture Radar (ASAR) instrument was an active radar sensor on-board the Envisat satellite see https://earth.esa.int/eogateway/instruments/asar

[41] SEASAT Data Processing Issues, 2014, https://earth.esa.int/eogateway/documents/20142/37627/SeaSat-SAR-IPF-Data-and-Processing-Issues.pdf

[42] What is the Freedom of Information Act? , UK Information Commissioner's Office, see https://ico.org.uk/for-organisations/guide-to-freedom-of-information/what-is-the-foi-act/

[43] What is eDiscovery? https://www.aiim.org/what-is-ediscovery

[44] Brownstone, R., Avoiding eDiscovery Nightmares: 10 Ways CEOs Can Sleep Easier, Forbes Magazine blog, 2011, https://www.forbes.com/sites/ciocentral/2011/06/15/avoiding-ediscovery-nightmares-10-ways-ceos-can-sleep-easier/

[45] *The chilling effect of the Freedom of Information Act 2000: How real is it?*, 2012, https://www.kingsleynapley.co.uk/insights/news/the-chilling-effect-of-the-freedom-of-information-act-2000-how-real-is-it

[46] Graham, C., Freedom of Information: this scaremongering leads to nothing but misinformation, the Guardian, 2012, https://www.theguardian.com/politics/2012/feb/19/freedom-of-information-scaremongering

moved to the computer system's "deleted" folder, erased from the folder itself, and then manually deleted from the backup system[47], [48], or access made difficult[49], [50].

2.8.2 Costs

As time passes more and more digitally encoded information is accumulated. It is therefore possible that the costs increase over time, yet experience tells us that the budget available for a preservation organisation usually does not. The following picture might therefore be projected to be the case.

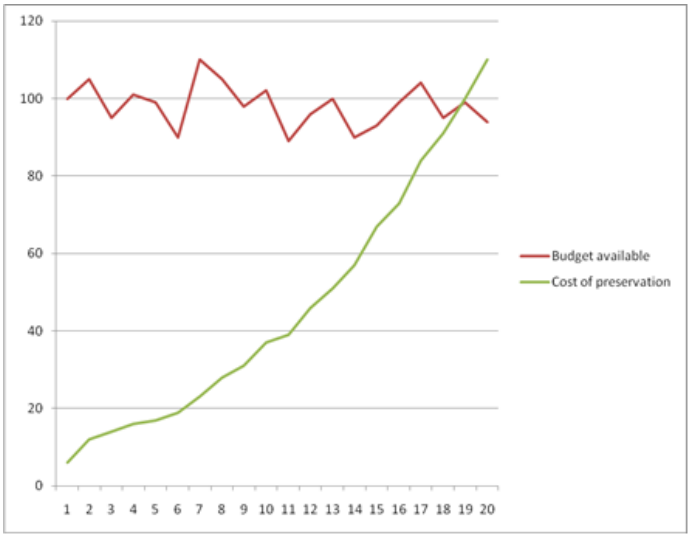


Figure 6 Money disincentives – if the annual cost of preservation of the accumulated data increases over time

If this is the projection then no responsible body would find it acceptable; a decision would have to be taken not to preserve everything – or perhaps not to preserve anything. The focus here is on how we could try to control the costs so that either the graph of preservation costs is level rather than increasing or is increasing only slowly so that the crossing-point is acceptably far into the future.

It is very hard to model the costs of digital preservation[51], [52], and even more difficult to evaluate possible benefits. However, it is worth discussing at least some of the costs at this point to illustrate the point.

---

[47] Report finds culture of "triple-delete" in B.C.. (n.d.) >*The Free Library.* (2014). Retrieved Sep 18, 2022, from https://www.thefreelibrary.com/Report+finds+culture+of+%27%22triple-delete%27%22+in+B.C.-a0457975535

[48] Proctor, J., CBS News, 2015, https://www.cbc.ca/news/canada/british-columbia/triple-delete-email-report-calls-for-penalties-for-foi-evaders-1.3367880

[49] Roberts, A.S. *Less Government, More Secrecy: Reinvention and the Weakening of Freedom of Information Law*, Public Administration Review, July/August 2000, Vol. 60, No.4

[50] Hood, C., *What happens when transparency meets blame-avoidance?*, 2007, see https://doi.org/10.1080/14719030701340275

[51] PADI web site with collection of information about costs of digital preservation https://webarchive.nla.gov.au/awa/20110824015945/http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/5.html

[52] Shehab, Essam & Lefort, Alice & Badawy, Mohamed & Baguley, Paul & Turner, Chris & Wilson, Michael & Conway, Esther. (2013). MODELLING LONG TERM DIGITAL PRESERVATION COSTS: A SCIENTIFIC DATA CASE STUDY. 10.13140/2.1.3837.5687.

One of the simplest costs which one may try to estimate is that of storage. The argument sometimes used is that the cost of a unit of storage reduces by a factor "r" each technology cycle (t_c years) for example 3 years. Suppose that the initial cost is £X. If at each cycle one buys new hardware then one spends £X/r in t_c years' time, £X/r^2 after a further t_c years, and so on. Therefore, one would spend

£X + £X/r + £X/r^2 + £ X/r^3 + ... £ X/r^n = £X(r-1/r^n)/(r-1) after "n" cycles, which, after several cycles is essentially £X(r)/(r-1) assuming r>1 . For example, if r=3 then the total cost over three or more cycles would be essentially 1.5*X.

Figure 7 shows how the multiplier approaches its final value after 3 or 4 cycles.

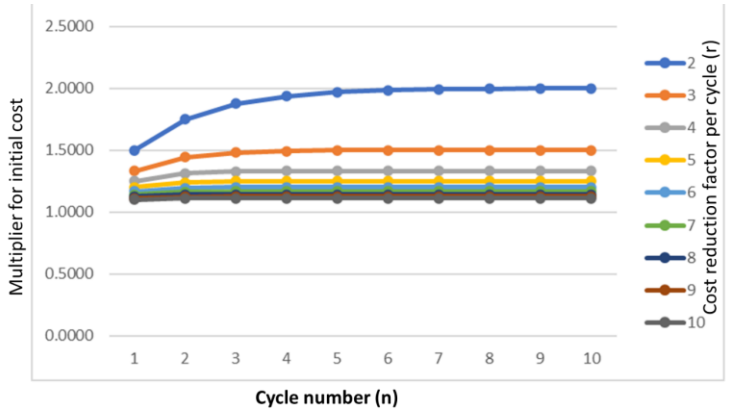


Figure 7 Change of multiplier over time (n) for different reduction factors (r)

Thus, one can argue that the hardware cost is at least controlled.

However, each cycle (for example 3 years) the amount of data may easily have increased by, say, a factor of m, thus the cost keeping all the data would be, after a few cycles:

Period 1		Period 2		Period 3			
X	+	X/r	+	X/ r^2	+...	=	r*X/(r-1)
		m*X/r	+	m*X/ r^2	+...	=	(m/r)*r*X/(r-1)
				m^2*X/r^3	+...	=	$(m/r)^2$*r*X/(r-1)
					+…		
						Total	(1+(m/r)+$(m/r)^2$+$(m/r)^3$….)*r*X/(r-1) which is [(r/m)/((r/m)-1)]*r*X/(r-1) assuming r>m

Table 10 Cost table – final value depending on

One can see that there is a real danger that the growth of data volumes may easily swamp the cost savings introduced by new technologies unless "m" is less than "r" i.e., unless the factor (m) by which the data grows in a certain period is less than the factor (r) by which the cost of the storage decreases. For example, if r=4 and m=2 then the total cost over very many years would be £8*X/3. Table 11 shows the value of the multiplier for values of r from 1 to 10, and for values of m from 1 to 10, for 10 cycles. When m is greater than r the sum becomes very large. When m=r the value is just the number of cycles, so the value grows linearly.

r→ m ↓	1	2	3	4	5	6	7	8	9	10
1	10.00	2.00	1.50	1.33	1.25	1.20	1.17	1.14	1.12	1.11
2	1023.00	10.00	2.95	2.00	1.67	1.50	1.40	1.33	1.29	1.25
3	29524.00	113.33	10.00	3.77	2.48	2.00	1.75	1.60	1.50	1.43
4	349525.00	1023.00	50.27	10.00	4.46	2.95	2.32	2.00	1.80	1.67
5	2441406.00	6357.16	246.57	33.25	10.00	5.03	3.38	2.64	2.24	2.00
6	12093235.00	29524.00	1023.00	113.33	25.96	10.00	5.50	3.77	2.95	2.48
7	47079208.00	110341.49	3587.06	357.85	69.81	22.03	10.00	5.90	4.14	3.24
8	153391689.00	349525.00	10909.75	1023.00	181.59	50.27	19.61	10.00	6.23	4.46
9	435848050.00	972874.83	29524.00	2659.41	445.06	113.33	39.70	17.98	10.00	6.51
10	1111111111.00	2441406.00	72578.52	6357.16	1023.00	246.57	80.27	33.25	16.81	10.00

Table 11 Multiplier for various m and r values for 10 cycles

However, this covers only the cost of preserving the bits; the cost of personnel; the cost of preserving the information have been left out of the calculations but may be estimated to grow linearly over time if no new types of data are added. In the simplest case where all the data is of the same type, the preservation activities will apply to all the holdings. In later sections we will see that there are three basic preservation activities:

- adding Representation Information needs to be done just once, to apply automatically to every object, but in any case this would not happen every cycle;
- transformations will require computational effort which scales with the volume of data, which grows by a factor m with each cycle, but there are mitigating factors:
  - one might expect the cost of each computation would decrease by a constant factor each cycle, reducing the growth of cost with time;
  - transformations would not be expected to happen every cycle;
- handover to another repository – requires the creation of AIPs which could be expensive to create but would happen just once at the end of the repository's custodianship.

More complex modelling is available based on cost data from real, anonymised, archives[53]. However, those cost models which are available omit the cost of maintaining understandability, which could be labour intensive. Large repositories are accumulating experience in cost modelling, and one would expect updated reports[54].

As a warning, I worked on one of these cost modelling projects (which I will not name), so it is my experience that some cost models seem to have been developed as a way to justify a specific amount of money, which is known to be available, for an archive (perhaps with some excess to allow funders to cut without harming the project, thus keeping all sides happy!).

The Blue-Ribbon Task Force on Sustainable Preservation and Access[55] has looked at the broader view and identifies the fact that, because the future is uncertain it is difficult to be

[53] Fontaine K, Hunolt G, Booth A, Banks M (2007) Observations on Cost Modelling and Performance Measurement of Long-Term Archives. Presented at PV 2007, Germany. Available from http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf

[54] Biscarat, C., Boccali, T., et al, 2020,New developments in cost modelling for the LHC computing, from https://www.epj-conferences.org/articles/epjconf/pdf/2020/21/epjconf_chep2020_03014.pdf

[55] NSF Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Web site http://brtf.sdsc.edu/

sure whether some object is worth preserving, therefore one should effectively purchase "future options" without making an indefinite commitment, as discussed in section 2.4.

2.9 Limitations of this chapter

There are many reasons to preserve, but what this chapter does not discuss is what is an essential determinant as to whether something is preserved or not, namely, how to secure that funding, but this is discussed in the next chapter.

3 Generating funding for preservation

3.1 Business Models for Digital Preservation

Section 0 provided reasons why digitally encoded information should be preserved. Some, such as legal requirements, will attract appropriate funding otherwise there will be dire consequences. Other sets of information can be exploited commercially, for example to attract eyeballs to advertisements. Value-added services may be provided using the preserved information, for example combining historical weather records with information about crops and harvest yield may be valuable to farmers.

Any particular exploitation route, or legal requirements, may only be applicable for a limited period, after which that specific business model will fail.

Another route must then be sought, or the surplus from the exploitation of some other holding must be used. As section 2.8.2 suggests, the cost of storage for the digital objects is likely to fall.

The APARSEN project examined many aspects of digital preservation, including the way in which value may be generated to fund preservation, illustrated in Figure 8.

Figure 8 APARSEN digital preservation value lifecycle

The activities may be summarised as follows – note that there are several forward references to chapters later on in this book, but because we are discussing these at a high level, the details can be looked at later:

- Preserve the object by a variety of sub-processes
  - Ingest selected information, based on the considerations of Chapter 0
  - Store – discussed in section 4.1
  - Plan preservation, including identifying the designated community (ideally this should be done at the earliest opportunity – certainly before the creation of the digital objects, if we want to secure the best conditions for future usage and we must secure a proper value justification to secure financial resources flows)
  - The basic preservation steps to counter changes are – as explained in in later sections:
    - create adequate Representation Information for the Designated Community and/or
    - transform to another file format if necessary or
    - if preservation cannot be carried on by the current organisation then hand over to the next organisation in the chain of preservation
  - Evidence about the authenticity of the digital objects must also be maintained, especially when the objects are transformed or handed over.

    - Confirmation of the quality of preservation can come from an Audit of the repository (with possible certification)
- Usability - discussed in detail in later sections.
    - Digital objects and digital collections should remain usable, i.e., one (human or artificial agent) should be able to understand and use the digital material. This is closely related to task performability. Various tasks can be identified and layered, e.g., rendering (for images), compiling and running (for software), getting the provenance and context (for datasets), etc. In every case task performability has various prerequisites, (e.g., operating system, tools, software libraries, parameters, representation information etc.). These prerequisites are termed Representation Information in OAIS, and the minimum amount of Representation Information needed is determined by the definition of the Designated Community.
    - Additional Representation Information may be created to enable a broader set of users to use and understand the digitally encoded information
        - Other communities may use different analysis tools, and it may be convenient to transform the digital object to a more convenient format. This will itself require its own Representation Information (RepInfo); the semantic RepInfo may be unchanged, but new structural RepInfo will certainly be needed.
    - The digital objects should also be discoverable in some sensible way – bearing in mind that some information will be publicly available whereas other information will be restricted.
- Value proposition – building on the discussions in chapter 0. The portfolio of Value propositions will provide the core of the answers to "Why preserve a certain digital collection and who would be willing to pay for it?"
    - Value propositions must be created by the identification, classification and quantification of the expected benefits which may be obtained by the targeted communities of customers and users from the continuous usage of the preserved objects, which in turn depends on the needs of the users and the usability conditions created for such preserved objects
    - the digital objects will probably be more useful to one type of user community than to another, and this may change over time. These differences and changes must be addressed by a portfolio of Value propositions (as well as by the design and implementation of adequate business models) chapter
    - rights may be associated with the digital objects, perhaps arising from the value or potential value of the object. These rights can generate revenue, and the revenue generation in turn depends on the business model used.
- Business case – building on the discussions in chapter 0.
    - There is an increasing demand from decision makers to justify: (1) the need for objects to be preserved, (2) the benefits derived of their usage, (3) the costs involved in the preservation, as well as (4) other resources required for preservation
    - One or more business models will address its implementation
    - There will almost certainly be options for trade-offs between costs, risks, and capabilities
- Business model
    - The business model lays out the business logic, i.e., how the value proposition is consistently delivered to the beneficiaries.

- Decisions about the mix of sources providing the financial resources required for implementing and operating the preservation business process will be based on the characteristics of the users and customers base (the target groups), the competition in the provision of the preserved assets as well as in the nature and dynamics of the formulated business case.
- The resources may be used at the very start to create new digital objects, which will presumably have been created for a specific purpose, and which then may be either disposed of or be preserved.
- A selection process will be needed to decide what is to be preserved. This will presumably be based on business case and risk considerations. It may also depend on the interest of other possible curators of the information.
- This financial resourcing may be (perhaps should be) part of the budgets needed to create the digital objects. However, some or all of the digital objects created may be disposed of rather than preserved.

3.2 Limitations of this chapter

This chapter has presented a way of looking fairly generally at justifying/generating the funding needed for preservation, but the specific benefits and business cases and models will vary enormously from one instance to another. They will depend upon the particular opportunities which present themselves together with the imagination and tenacity of the reader.

4 Diving into the BITS in question

This chapter explores why it is difficult to preserve information which is encoded in digital objects. In order to do this, one must look at a selection of digital objects starting at the bit level, showing how these bits must be taken together in order to eventually be understood. Enroute it will become clear that there are many things' people do not normally consider simply because of the support, often behind the scenes and therefore effectively invisible, of software which is commonly available.

4.1 Keeping the bits safe

The obvious point to be made is that one cannot normally see the bits and bytes. One normally needs special hardware/software, which can become unavailable. However, this is not always the case.

Bits can be kept on diverse types of media. For example, consider the "1"s and "0"s could be carved in stone – easily viewable for the next thousand years, and so at least in this case one can see the bits. Similarly, one could print the "1"s and "0"s on paper or film, which can last for many decades.

Sometimes what one can see and may think are the bits may not be. Consider an old-fashioned CD-ROM[56]. There are (re-)writable CD-ROMs but a mass-produced CD-ROM is made by stamping small pits into a round plastic blank.

With a microscope one can see something like the image in Figure 9. But even though one sees little pits in the material, these are not the bits. There is a great deal of error correction[57] so that errors in the pits are corrected by taking the raw signal from the pits in the disk and processing it in the electronics to produce the actual bits.

Figure 9 CD-ROM showing data pits

The technology of the hardware and software of the readers changes, sometimes very quickly.

There are many types of media which still exist but for which the readers are obsolete[58] , not repairable or simply not available. Hardware often requires firmware in order to work correctly. Software is also needed to get something that can be processed.

Because one is normally so far removed from the bits, another question is how one can tell that the bits have not been changed. As noted above, the underlying physical structure, whether pits on a plastic disk or magnetic domains or the state of an electronic switch, can have errors and the redundant "bits" are added which the reader (hardware/firmware/software) can use to correct certain types of errors automatically, before they are presented to a user. A special case is where disks are combined in RAID systems in which the redundant bits are spread between multiple disks.

---

[56] See https://www.britannica.com/technology/CD-ROM

[57] https://www.irishtimes.com/news/science/high-fidelity-how-the-sound-of-cds-stays-error-free-1.2362987

[58] The museum of obsolete media https://obsoletemedia.org/

But what happens if an error occurs which cannot be corrected, or even worse, cannot be detected automatically? What happens if the bits have been altered by some outside agency?

Should we worry about such things? A detailed report was written some years ago[59] but which provides a valuable collection of ideas.

The rate of uncorrectable bit errors, per transmission or over a certain time, is tiny (figures of 10^{-16} or smaller) but if the number of bits is large then the likelihood of silent errors occurring will be large. At the time of writing people think little of filling a 1TB disk whether attached to their computer or offered "free" by Google or Facebook, and scientific projects talk in terms of needed to store 10s, perhaps 100s of Petabytes (1 Petabyte is 2^{50}, which is about 10^{15} bytes), with larger sizes – Exabyte (2^{60}), Zettabyte (2^{70}) and Yottabyte (2^{80}) - being discussed. Several commercial systems[60] are available which can cope with many 100s of Petabytes and can certainly scale further..

A way to detect whether a digital object has changed is to keep several copies and then do bit for bit comparisons. If they are all different from each other then we are in trouble but if one can find a majority which are the same then the odd one out has changed and must be corrected – by replacing it by one of the good objects. However, keeping lots of copies (let's say N copies) and doing a bit-by-bit comparison between them all, which would need N(N-1)/2 comparisons, which is very expensive.

The commonly used alternative is to calculate a hash for the digital object. A hash is fixed length digital object which is calculated by applying an algorithm to the digital object of interest. The point is that[61]

> *The same message always results in the same hash. Ideally it should also have the following properties:*
> - *it is quick to compute the hash value for any given message*
> - *it is infeasible to generate a message that yields a given hash value (i.e., to reverse the process that generated the given hash value)*
> - *it is infeasible to find two different messages with the same hash value*
> - *a small change to a message should change the hash value so extensively that a new hash value appears uncorrelated with the old hash value*

There are several different ways to calculate a hash, each one results in a different hash value[62].

Having chosen one specific hash algorithm one then calculates the hash for the digital object and keeps that hash value very safe. Then in the future one can recalculate the hash and if this value is identical with the one which we have kept safe then one be pretty sure that the digital object has not changed. Of course, there is a chance that, by accident two different objects will have the same hash, simply because there are only 4,294,967,296 different values of a hash with 32 bits. If we calculate hashes for more than 10 billion objects, then there is a very good chance that two of the hashes will be the same. Therefore, one normally calculates multiple hashes, each with a different algorithm, in order to reduce the chance of an error – if any of the recalculated hashes are different then the object must have changed.

---

[59] Rosenthal, D.S.H., 2010 Keeping Bits Safe: How Hard Can It Be? available from https://queue.acm.org/detail.cfm?id=1866298

[60] For example, see LABDRIVE from Libnova https://docs.libnova.com/labdrive/, and in particular https://docs.libnova.com/labdrive/concepts/architecture#storage

[61] See https://en.wikipedia.org/wiki/Cryptographic_hash_function

[62] Kiao, U, *Probability of Collision in Hash Function*, see https://iq.opengenus.org/probability-of-collision-in-hash/

To keep a hash value safe is therefore very important. It is just another digital object but since it is much smaller than the object for which the hash was created, a very large number of copies can be kept, as described above. The hash could even be printed in a newspaper advertisement, or it could be carved in stone! Alternatively, a hash of multiple hashes could be calculated, and that final hash could be saved[63].

The hashes would need to be calculated periodically because changes/bit rot/corruption of a digital object could happen at any time. The practicalities of calculating hashes can be challenging, and costly, in terms of the computing power needed when the number and volume of objects are huge. Cloud providers enable the compute resources to be available only when needed, which is made easier by software such as Kubernetes[64]. The time taken to calculate the full set of hashes will limit the frequency of re-calculating the hashes.

If a digital object is found to have changed, if that is the only copy then we have then we are in trouble, because there is no way to go backwards from the hash to reconstruct the digital object. Therefore, one would need to have kept at least one other copy and hope that that has not changed. Three or more copies would be safer. If one recalculates the hashes very infrequently then one should have more copies because there is a greater risk of random changes the longer time passes.

On the other hand, purposeful corruption may affect all copies simultaneously. This may seem ridiculous but if the bits are valuable to a large company or even a nation, then vast resources could be devoted to that purposeful corruption.

4.2 What do the bits mean?

Perhaps the reader has never bothered to look at the bits inside a digital object – why would you?

In this section I hope to persuade you that it is an important part of your mental toolkit for thinking about digital preservation.

Well, it is true that one can drive a car without knowing how the vehicle works. But what happens when it breaks down. Someone needs to know how to fix it and, to do that, someone needs to know how it works. In the case of a car, one could simply dump it and get another one.

But if we are thinking about digital preservation then in the future if one cannot understand the digital object there is no analogy of buying a new car! In other words, if the readers intend to preserve digitally encoded information then they must understand what is "under the hood" of digital objects.

A simple example is the following sequence of bits:

01000001 01010100……..

What does it mean? If we have spent a great deal of effort keeping this safe for perhaps centuries but we don't know what it means or is used for then surely we have failed in the task of preserving the digitally encoded information.

[63] Smorul, M., Song, S., JaJa, J.,2009, ACE: a Novel Software Platform to Ensure the Long-Term Integrity of Digital Archives, available from https://library.imaging.org/admin/apis/public/api/sandbox/website/downloadArticle/archiving/4/1/art00022

[64] Kubernetes, also known as K8s, is an open-source system for automating deployment, scaling, and management of containerized applications, see https://kubernetes.io/

On the other hand, if we are told that this is encoded in 7-bit ASCII, and we have the following table, perhaps on a piece of paper, or carved in stone or etched on a titanium sheet, we have a chance as will be described next.

BIN	HEX	Symbol	BIN	HEX	Symbol	BIN	HEX	Symbol	BIN	HEX	Symbol
0000000	0	NUL	0100000	20		1000000	40	@	1100000	60	`
0000001	1	SOH	0100001	21	!	1000001	41	A	1100001	61	a
0000010	2	STX	0100010	22	"	1000010	42	B	1100010	62	b
0000011	3	ETX	0100011	23	#	1000011	43	C	1100011	63	c
0000100	4	EOT	0100100	24	$	1000100	44	D	1100100	64	d
0000101	5	ENQ	0100101	25	%	1000101	45	E	1100101	65	e
0000110	6	ACK	0100110	26	&	1000110	46	F	1100110	66	f
0000111	7	BEL	0100111	27	'	1000111	47	G	1100111	67	g
0001000	8	BS	0101000	28	(	1001000	48	H	1101000	68	h
0001001	9	HT	0101001	29	)	1001001	49	I	1101001	69	i
0001010	0A	LF	0101010	2A	*	1001010	4A	J	1101010	6A	j
0001011	0B	VT	0101011	2B	+	1001011	4B	K	1101011	6B	k
0001100	0C	FF	0101100	2C	,	1001100	4C	L	1101100	6C	l
0001101	0D	CR	0101101	2D	-	1001101	4D	M	1101101	6D	m
0001110	0E	SO	0101110	2E	.	1001110	4E	N	1101110	6E	n
0001111	0F	SI	0101111	2F	/	1001111	4F	O	1101111	6F	o
0010000	10	DLE	0110000	30	0	1010000	50	P	1110000	70	p
0010001	11	DC1	0110001	31	1	1010001	51	Q	1110001	71	q
0010010	12	DC2	0110010	32	2	1010010	52	R	1110010	72	r
0010011	13	DC3	0110011	33	3	1010011	53	S	1110011	73	s
0010100	14	DC4	0110100	34	4	1010100	54	T	1110100	74	t
0010101	15	NAK	0110101	35	5	1010101	55	U	1110101	75	u
0010110	16	SYN	0110110	36	6	1010110	56	V	1110110	76	v
0010111	17	ETB	0110111	37	7	1010111	57	W	1110111	77	w
0011000	18	CAN	0111000	38	8	1011000	58	X	1111000	78	x
0011001	19	EM	0111001	39	9	1011001	59	Y	1111001	79	y
0011010	1A	SUB	0111010	3A	:	1011010	5A	Z	1111010	7A	z
0011011	1B	ESC	0111011	3B	;	1011011	5B	[	1111011	7B	{
0011100	1C	FS	0111100	3C	<	1011100	5C	\	1111100	7C	|
0011101	1D	GS	0111101	3D	=	1011101	5D	]	1111101	7D	}
0011110	1E	RS	0111110	3E	>	1011110	5E	^	1111110	7E	~
0011111	1F	US	0111111	3F	?	1011111	5F	_	1111111	7F	

Figure 10 7-bit ASCII codes - binary/hexadecimal/character

In a “byte” there are 8 bits. Here we are looking at the case where the first bit is “0”, but that is not shown; there will be another table the same size for the bytes which start with “1”, but we don’t need to look at those now.

In this diagram the 128 different characters are shown here in 4 groups each of 3 columns. The column labelled BIN shows the 7 bits used in the encoding.

The column labelled HEX shows a convenient way of showing that bit sequence. The 8 bits are split into 2 groups of 4 (remembering that we are not showing the leftmost “0”. 4 bits can present 16 different vales (hence the name **hex**adecimal). These are shown as 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F

Any sequence of BYTES is a sequence of BITS which can be written as “1”s and “0”s – and this can be written in a short form in HEX. So, for example you see that “01000001” if divided as “0100” and “0001”, which in hexadecimal are “4” and “1” so this is written as “41”, which makes it much easier to read and compare.

In ASCII hexadecimal “41” represents symbol “A”, as can be seen in the figure above.

Similarly, BITS “01010100” is HEX 54 which can be seen in the table represents “T”

Therefore, given the code table we know that the bits represent text that begins “AT” – clearly we can decode the remaining characters in the same way and get the full text.

Let us look next at the following bit sequence

01001110 01001101 01010001 01001101 01010000 01001010 00100000 00100000

These bits are written in HEX as

4e 4d 51 4d 50 4a 20 20

which makes it much easier to "read."

If this is something encoded in ASCII-7 then it would represent:

NMQMPJ

There are many possible encodings, for example the IEEE 754 encoding for 32-bit real numbers, in which case the bits would represent the following two real numbers:

$8.6116461*10^{8}$ and $1.35644119*10^{10}$

It could also represent two 32-bit integers:

164211241 and 168379396

There are many, many, other possible encodings. Each bit may mean something different. For example, if the first bit is "1" may mean "Make some coffee", but if the first bit is "0" it may mean "Turn the kettle on and make some tea"; then if bit number 2 is "1" it means "add milk if the first bit is 1 but add lemon if the first bit is 0". In other words, there may be many dependencies such that the meaning of the bits changes depending on the other bits.

In fact, I know that, in this specific case, the answer is NMQMPJ – because I know the encoding is 7-bit ASCII and the meaning – the **semantics** - are that this was my flight reference some years ago and was quite important for me at that time.

Here is another example with some scientific data – how it is encoded does not matter – as we have seen there are many different encodings possible. What we know is that the bits represent the following text.

longitude	latitude	Ozone	Date
132	50	34.9	12/03/1999T17:20:43.1
178	50	45	12/03/1999T19:37:52.7
190	50	78	12/03/1999T21:16:23.9

Table 12 Example table

We want to focus here on semantics. We can see that the data is basically a table with columns – we can see the names of the columns. But what do they mean? We consider each column in turn.

latitude/longitude – taken as a pair we can guess that this gives a location and looks as if the units are "degrees." But is it latitude/longitude on EARTH? Or is it VENUS? or the MOON? In fact, it is data about the Earth – but it need not have been! Also, there are other latitude/longitudes – for example MAGNETIC

Next the **OZONE** column. Is it a measurement like the density of Ozone at a point somewhere above the location? Or it could be an integration of all the Ozone along the line of sight. Or it might not have anything to do with the gas called Ozone, instead it might be an acronym for something entirely different.

Finally, **TIME** – what time zone? daylight saving? Does 12/03/1999 mean 12th March 1999, as a European would think, or is it 3rd December 1999, as it would be understood in the USA? There are many other possibilities, refinements, and questions we could go into.

However, this brief view shows that, even if all the things are common knowledge to the creators of the table, in the future it will not be obvious, and indeed the table will be essentially useless without knowing the semantics associated with the columns.

4.3 Another complication with the bits

Consider a simple example of CSV[65] files containing the text:

FirstName, Surname, Gender Fred,Bloggs, Male Jane, Bloggs, Female

The files may be saved, for example using the Windows Notepad application using UTF-8, UTF-8 with Byte Order Mark[66] (BOM), UTF-16, and a variety of others. **Current** applications such as Excel or Notepad, are likely to deal with these in an apparently identical fashion. However, the bit sequences are different.

Encoding	Hex first line
UTF-8	46 69 72 73 74 4e 61 6d 65 2c 20 53 75 72 6e 61
UTF-8 BOM	ef bb bf 46 69 72 73 74 4e 61 6d 65 2c 20 53 75
UTF-16	ff fe 46 00 69 00 72 00 73 00 74 00 4e 00 61 00

In the future, applications may not be so accommodating, and may not recognize one or other of these encodings automatically

The PRONOM[67] code for **all** these is **x-fmt/18** and the PRONOM page further tells us that the MIME Type is **text/csv** but provides no further information about encoding.

More importantly no information about the semantics is provided. In this case it may be obvious what *FirstName*, *Surname* and *Gender* "mean", but does the latter mean gender at birth or by declaration or following medical procedure?

The software which may be used to deal with these files is readily available now, for example Excel or Notepad in Windows, but will they be so readily available in future? Moreover, while UTF-8 and UTF-8 BOM appear using Excel as a table with 3 columns, the UTF-16 is shown as having only 1 column.

This is a trivial example, and one can imagine the difficulties which can arise for more complex scientific and other data, such as that belonging to scientific research organizations.

Scientific data may use specialised terminology. The current users of that data will be very familiar with terms such as

- bad_thing/min_x = -390
- coffset = 582.00e-3
- SUBRUN

Such current users will understand the meaning, units, and use of such terms. In the future such things may not be common knowledge when a project has ended but the data is still used. In some cases, it may be possible to guess the meaning, but guesses may be catastrophically wrong.

[65] See https://docs.fileformat.com/spreadsheet/csv/

[66] See https://www.w3.org/International/questions/qa-byte-order-mark

[67] https://www.nationalarchives.gov.uk/PRONOM/

4.4 A deeper dive into documents

Here we look in detail at a variety of document files so that we can understand some of the complexities which we normally ignore.

The files we will consider were each created with the text

"This is a test"

The files have names where the file extension indicates the file type, which is the norm with an MS Windows operating system, but NOT if one is using a UNIX-type operating system such as Linux[68].

Test.doc	The "old" MS Word format
Test.docm	MS Word which contains macros
Test.docx	The new MS word file
Test.odt	OpenOffice Document file
Test.pdf	A PDF file
Test.rtf	A Rich Text Format file
Test.txt	An ASCII text file

Table 13 Test file names

The following shows screen shots from the Treeview[69] application, first for the Test.txt file.

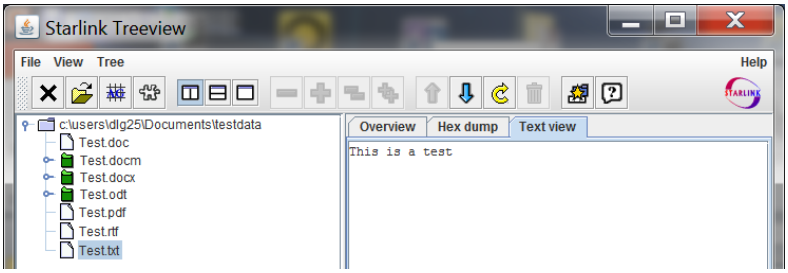


Figure 11 Internals of "Test.txt"

One can inspect the Hexadecimal in the following diagram.

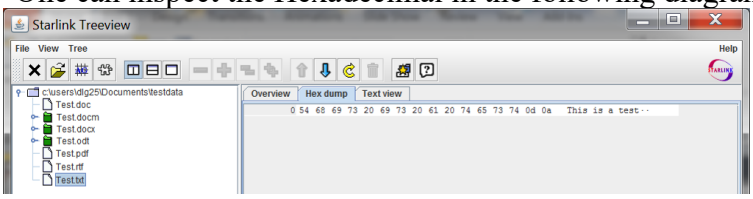


Figure 12 Looking at "Test.txt" as Hexadecimal

One can see the "54" "68" "69" which is "T" "h" "i". The full "This is a test" is shown to the right of the hexadecimal.

[68] See https://www.linux.org/ and https://www.linuxfoundation.org/

[69] http://www.star.bristol.ac.uk/~mbt/treeview/

On a Linux computer one can use the command “xxd,” as will be illustrated below.

On a Linux computer one can use the following command to show the first 120 bytes of test.txt with 16 bytes per line

```
xxd -l 120 -c 16 test.txt
```

which produces:

```
00000000: 5468 6973 2069 7320 6120 7465 7374 0d0a  This is a test..
```

which matches the output one can see above.

Figure 13 Using Linux command line to display hexadecimal

Looking next at the .rtf file:

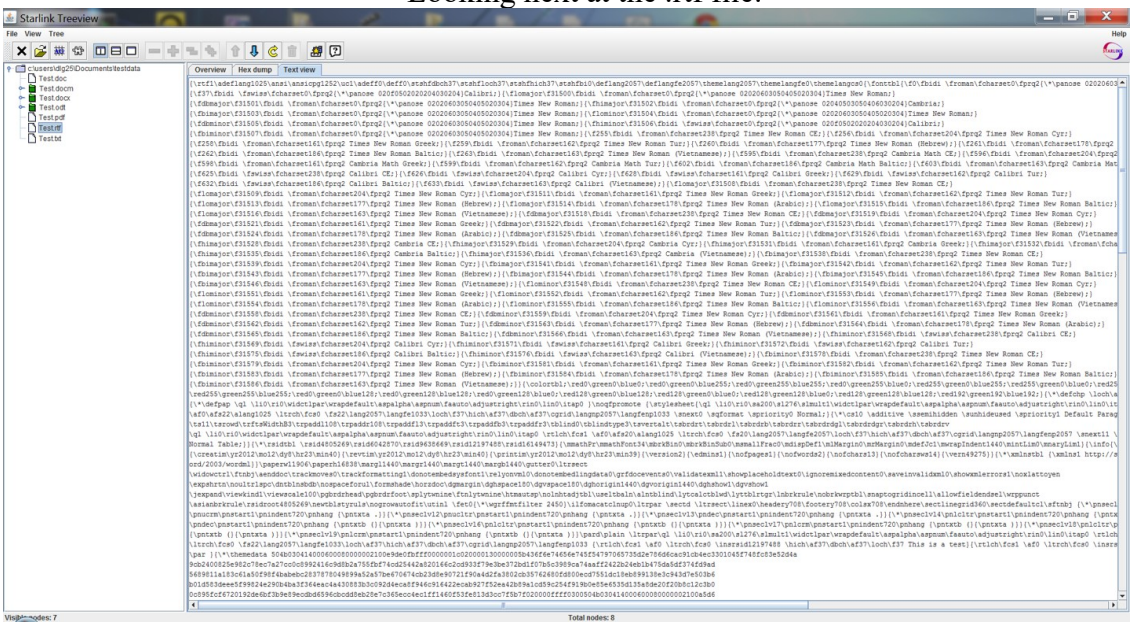


Figure 14 Looking at "Test.rtf" as Hexadecimal

Six lines from the bottom one can see “This is a test.” The rest of the file is presumably to do with displaying the text. At the start of the file, one sees “\rtf1” which is an indication that this is an RTF file.

Next we can examine the PDF file:

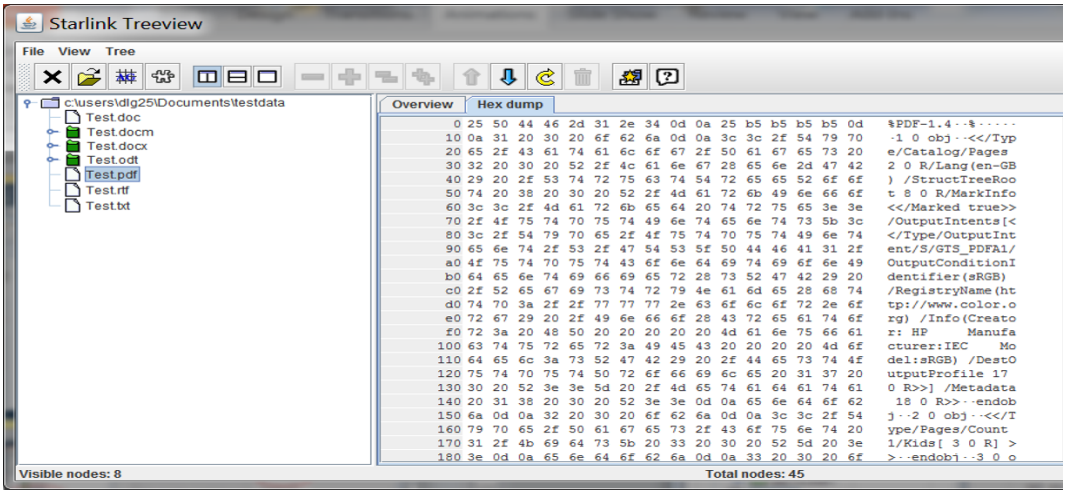


Figure 15 Looking at "Test.pdf" as Hexadecimal

The text “This is a test” is much further down in the file and cannot be seen in this figure. However, one can see that at the start of the file there is “%PDF-1.4” which indicates that this is PDF version 1.4 file, which we will come back to later.

The ODT file is a little different:

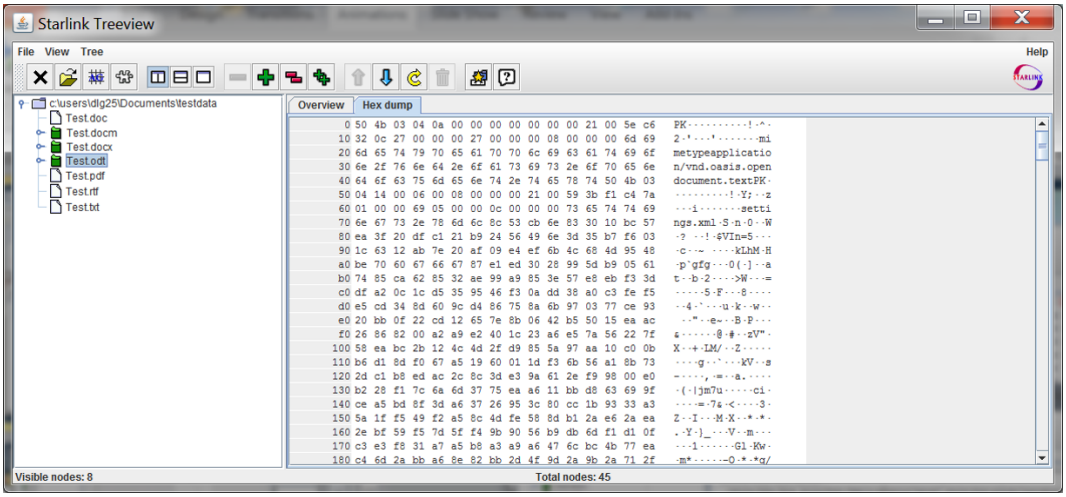


Figure 16 Looking at "Test.odt" as Hexadecimal

This is a file which must be unzipped – as suggested by the starting characters “PK.” Treeview unzips the file internally and one sees:

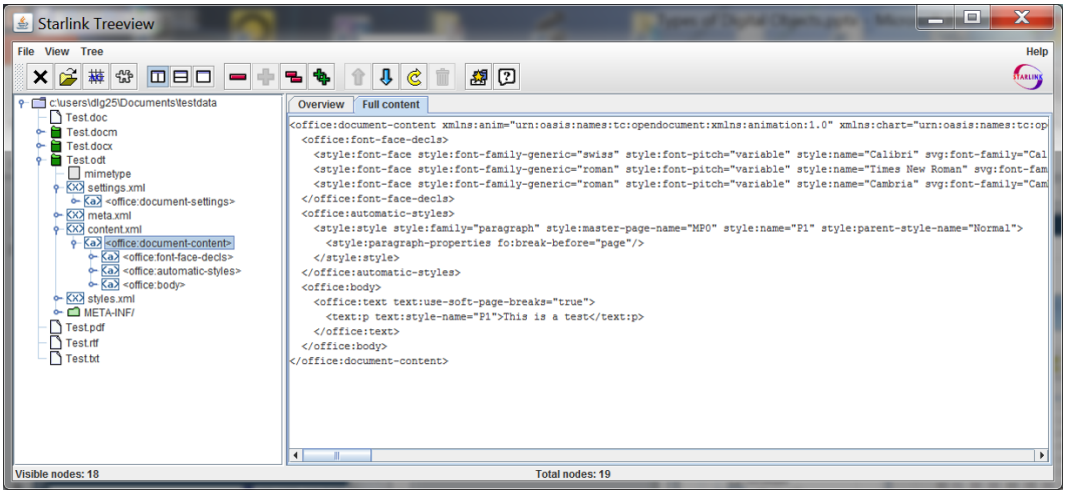


Figure 17 Looking at uncompressed "Test.odt" as Hexadecimal

This shows that the ODT, once unzipped, is a set of folders and XML files. Inside of which a component “office document content” shows the text “This is a test.” The remainder is presumably to do with displaying the text. It is worth noting that there are references such as

“urn:oasis:names:tc:opendocument:xmlns:animation:1.0”

This seems to be a reference to something on the internet -perhaps. The definition of the Open Document Format for Office Applications[70] declares an XML Namespace with this name.

Let us look now at the MS Word formats, starting with the “Test.doc” file:

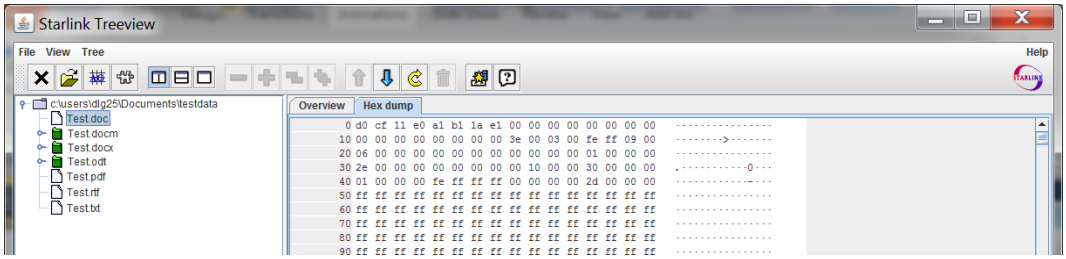


Figure 18 Looking at "Test.doc" as Hexadecimal

We see that most of it does not print as ASCII characters. But moving further down the file we see the following.

[70] http://docs.oasis-open.org/office/v1.2/cs01/OpenDocument-v1.2-cs01-part1.html

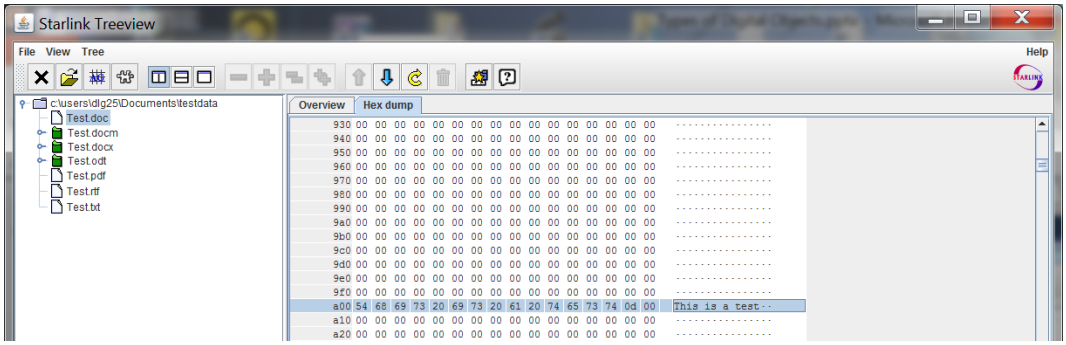


Figure 19 Looking further down in "Test.doc" as Hexadecimal and text

Here we see the text “This is a test.” The rest of the bytes look as if they are blanks e.g., HEX “00” or “ff”, which may be placeholders. Presumably, the formatting instructions are elsewhere in the file.

We turn now to the “Test.docx” file – the newer MS Word format.

You may not be surprised to learn that this is also a compressed file, as with the “Test.odt” file. The reader can see this for themselves by changing the “.docx” filename extension to “.zip.” This can then be “unzipped” in the normal way. One will then see the internal structure, as shown in the following figure.

Indicated in the red ellipses we see that in “document.xml” there is the text “This is a test.”

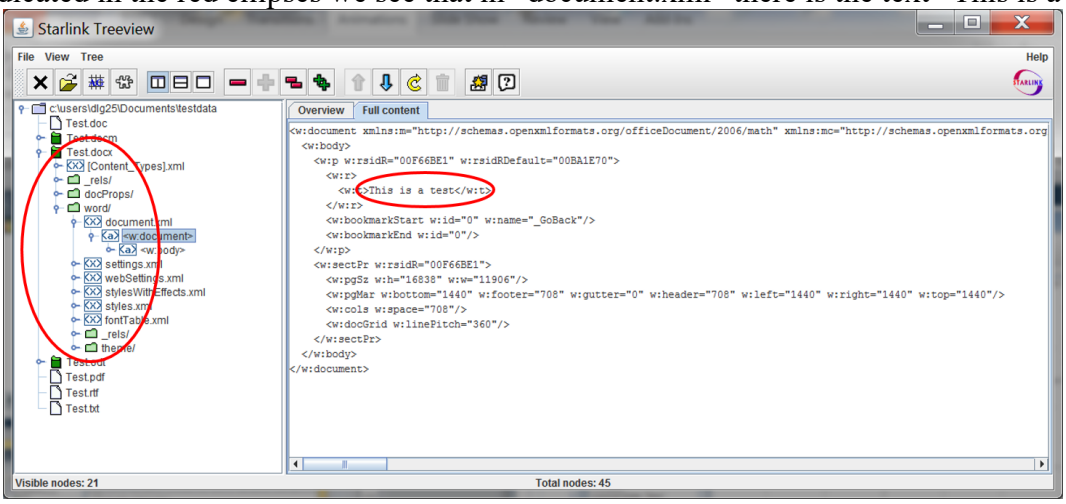


Figure 20 Looking at uncompressed "Test.docx" as XML text

There are also namespaces defined using URI

http://schemas.openxmlformats.org/officeDocument/2006/math

One can find some details of this namespace at http://www.datypic.com/sc/ooxml/ns-m.html but one wonders whether it is important that the application can access things on the Internet. Of course, Word still works when your computer is not connected to the network but maybe it caches (i.e., stores on a local disk) whatever information is needed but usually such cached systems check with the original source periodically. What will happen in the future if such resources are no longer available?

Maybe everything will work fine in future, even if one has the MS Word application running on an emulator, but we would be in trouble if we rely on MS Word, and something goes wrong.

Looking back at the examples here we can see that these document formats have a lot going on that we don't understand. We rely on the software, which is currently readily available, from Web browsers to MS Word or Adobe Acrobat reader.

4.5 A deeper dive into tables

There are some more useful things we can glean by looking at a number of ways in which a spreadsheet can be stored on this. As was done with documents, we take the information in Table 12 and write it out in a variety of formats, then use Treeview to examine the details.

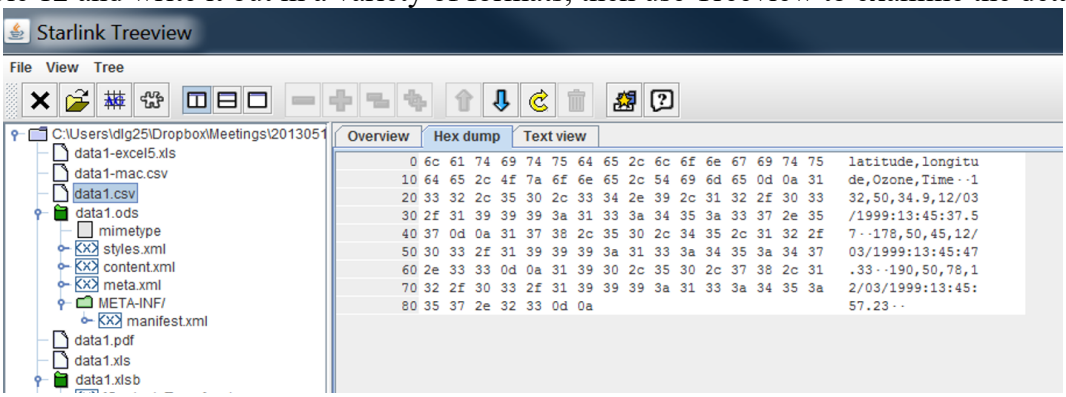


Figure 21 Looking at "data1.csv" as Hexadecimal and text

The first example is a ".csv" file, which is simply a text file where the value of each column is separated from the adjacent columns by a comma. Of course, if the value itself contains a comma (,) then the value is normally enclosed by inverted commas (").

BIN	HEX	Symbol
0000000	0	NUL
0000001	1	SOH
0000010	2	STX
0000011	3	ETX
0000100	4	EOT
0000101	5	ENQ
0000110	6	ACK
0000111	7	BEL
0001000	8	BS
0001001	9	HT
0001010	0A	LF
0001011	0B	VT
0001100	0C	FF
0001101	0D	CR

Figure 22 Extract from ASCII-7 table

Looking at the table of ASCII characters we have the hexadecimal "6c" which is "l", "61" which is "a" , and so forth. That all looks rights. We mentioned in the earlier section about the BOM but that does not seem to be here.

However, when we look at the end of the first line – we can locate this because "Time" is hexadecimal "54,69,6d,65" and this is followed by "0d,0a". Looking at the ASCII table we see "0d" represents CR - "Carriage Return" and "0a" represents LF - "Line Feed". Those readers who have seen or used old manual typewriters for US or European users will understand what these terms mean. The typist when coming to the end of a line used the handle at the right-hand side to push the carriage, which holds the paper, to the left – the Carriage Return. Then the handle was given another push, and the paper was rolled up by one line – Line Feed. The CR/LR was used when sending messages to old teletypes to print out text.

The CR/LF end of line is what one gets with the Microsoft operating system applications.

We can look at what one gets with a MacIntosh:

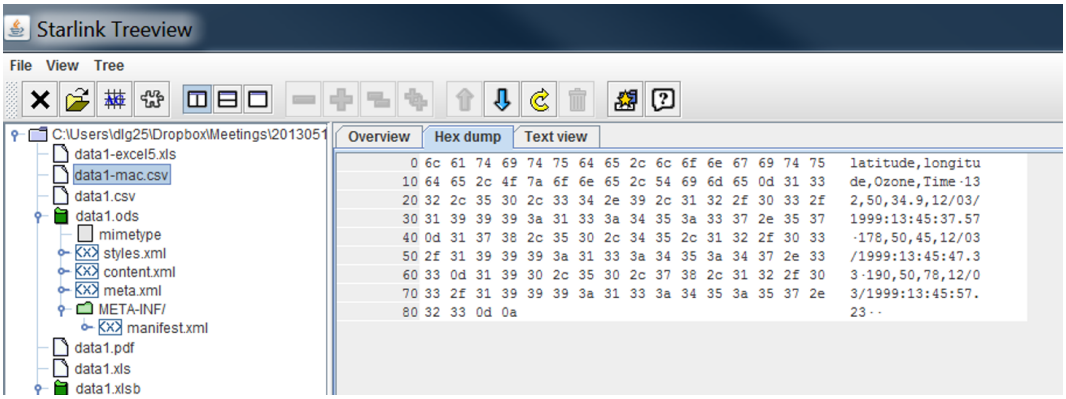


Figure 23 Looking at "data1-mac.csv" as Hexadecimal and text

Here one sees that the end of line is marked with “0d” – i.e., only “Carriage Return”.

Currently when one copies a file from a MacIntosh to a Windows computer the operating systems “automagically” makes the adjustment.

Turning now to Excel using the new format “.xlsx” we see:

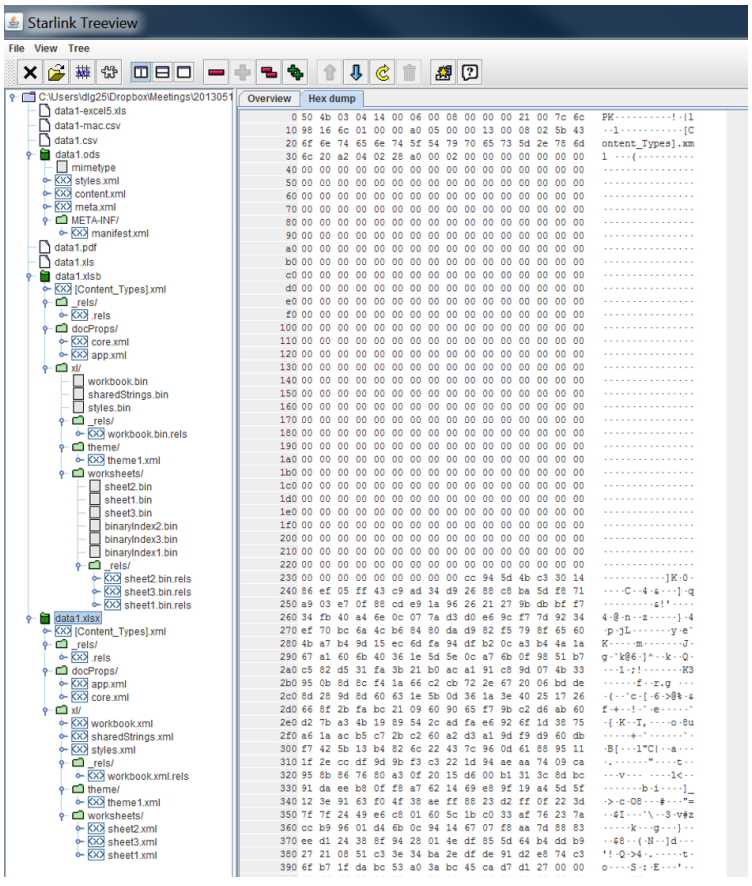


Figure 24 Looking at "data1.xlsx" as Hexadecimal

As with the “docx” files we see the file starts with the text “PK” which is a hint that the file is compressed, and so, as with “docx” we can unzip it, which the Treeview application does automatically:

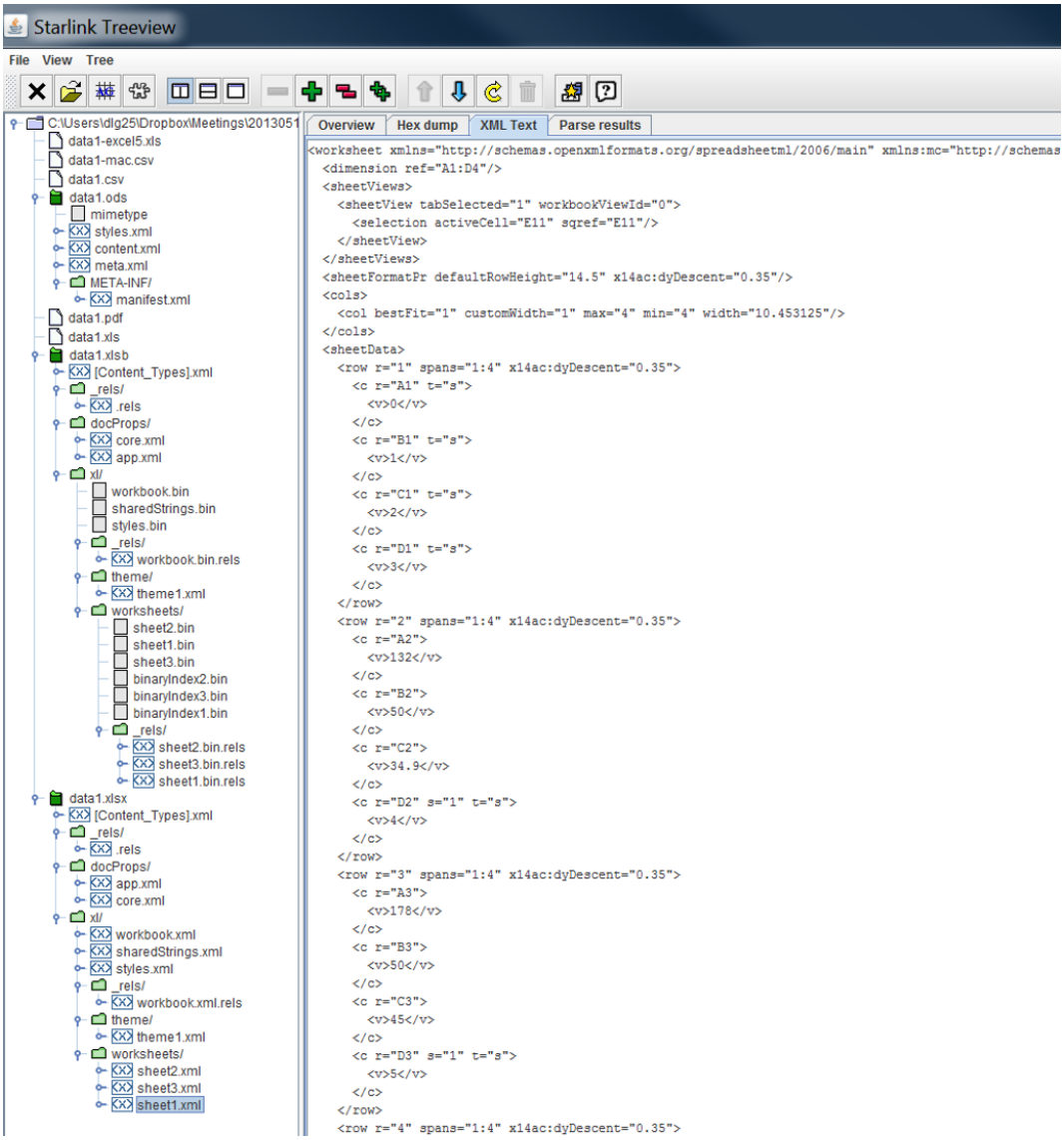


Figure 25 Looking at uncompressed "data1.xlsx" as XML text

Examining “sheet1” we see the numbers. As with “docx” there are namespaces and a large number of things we do not recognise, but which the software the is available at the time of writing, takes care of automatically.

We can do the same thing with an Open Document spreadsheet “.ods” and recognise the column headers and some of the numbers in “content.xml.”

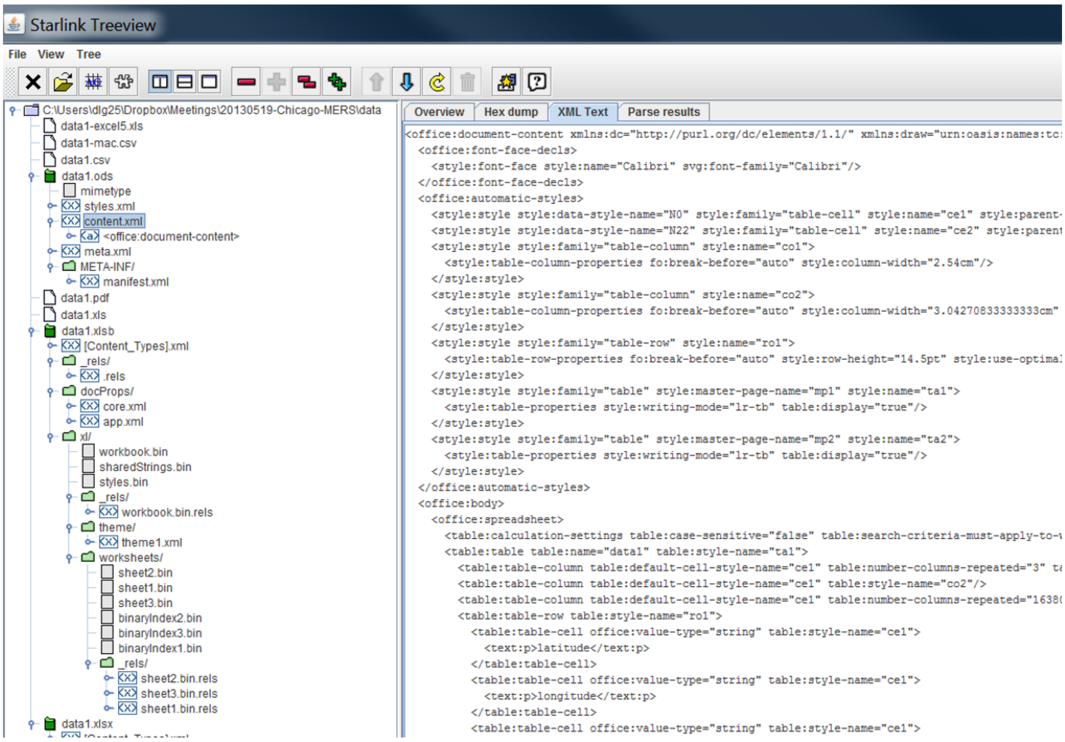
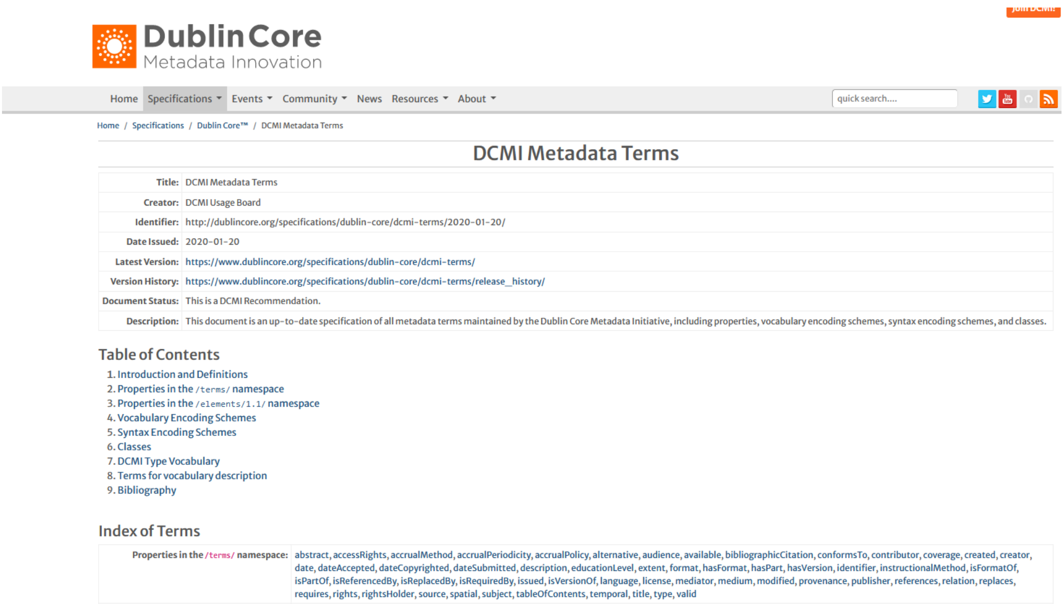


Figure 26 Looking at uncompressed "data1.ods" as XML text

We can also see URIs such as http://purl.org/dc/elements/1.1 - this is a URL which, at the time of writing, does resolve to:



But will this web page always be there, and what happens if/when it is not?

4.6 Images, Audio and Video

There are of course many other types of digitally encoded information, including images, audio, and video; these are examined, using Treeview, next.

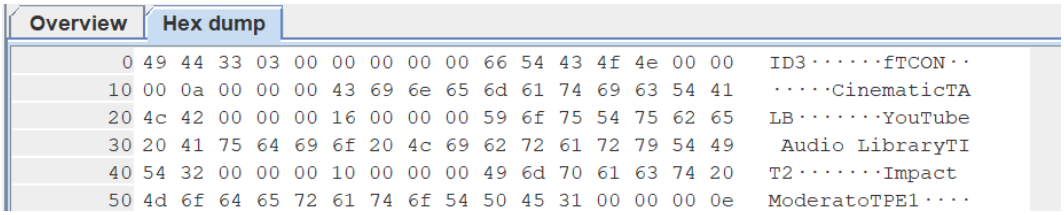


Figure 27 MP3 audio file

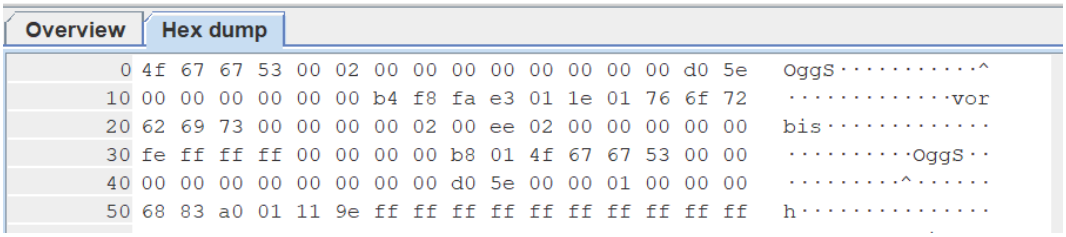


Figure 28 OGG audio file

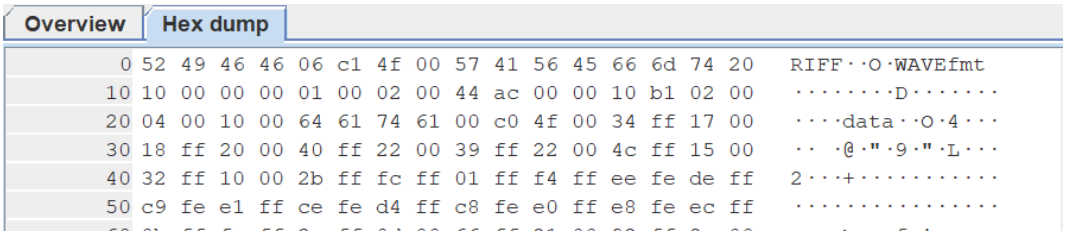


Figure 29 WAV audio file

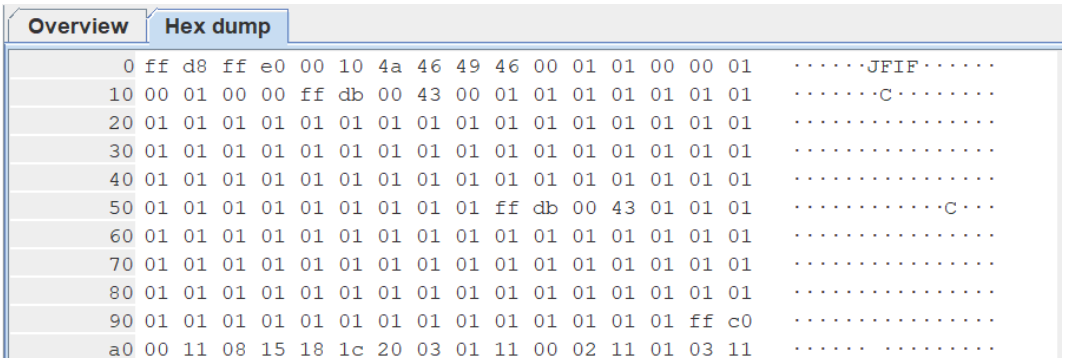


Figure 30 JPG image file

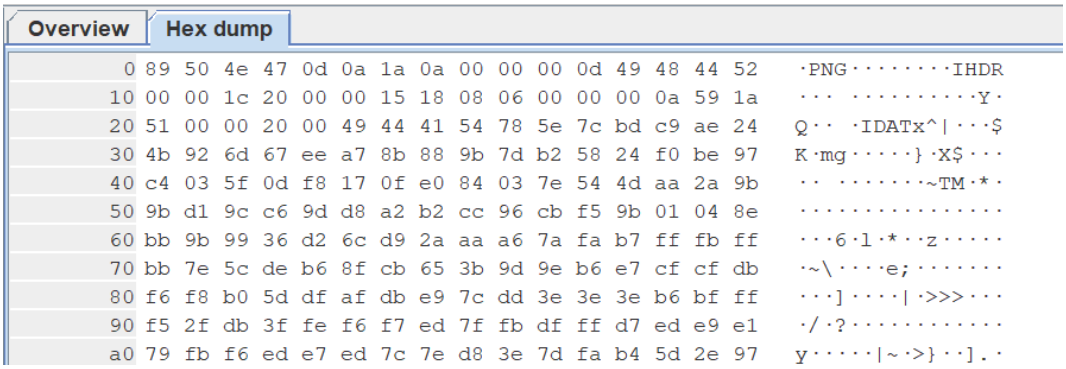


Figure 31 PNG image file

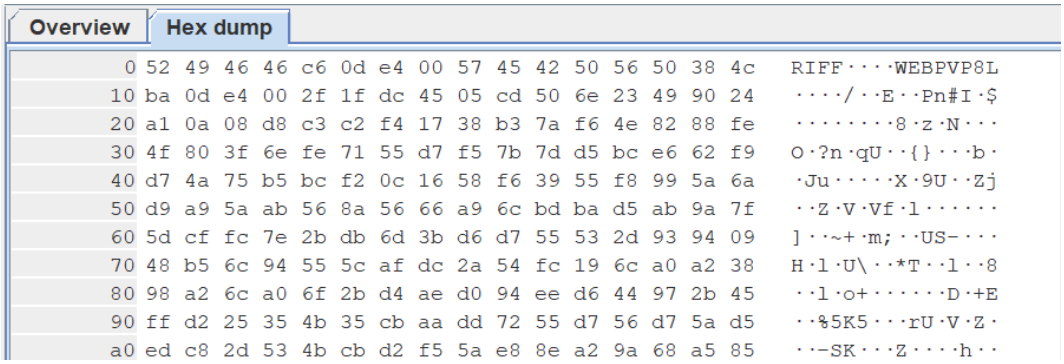


Figure 32 WEBP image file

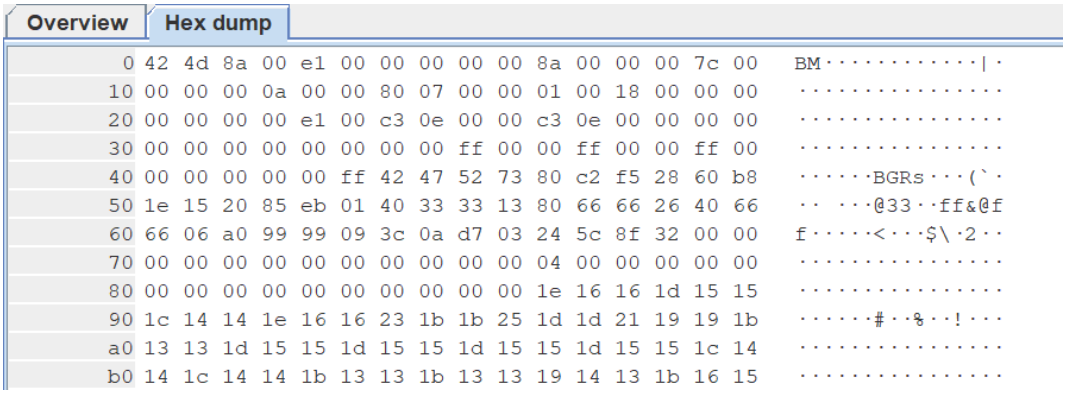


Figure 33 BMP image file

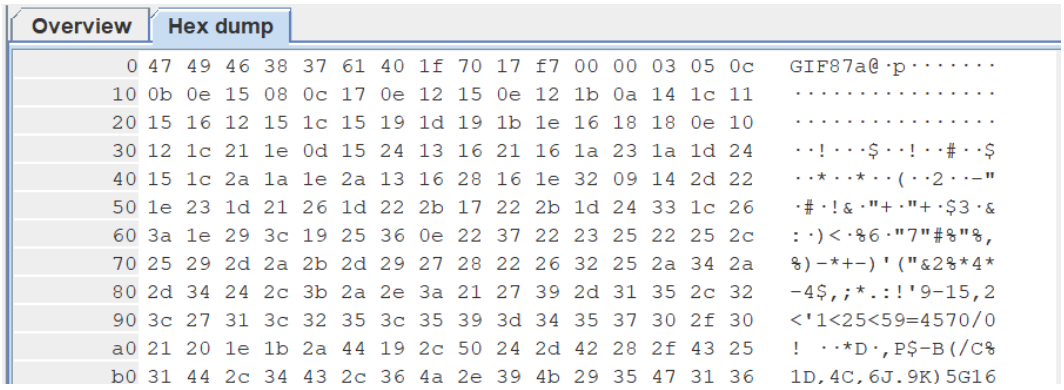


Figure 34 GIF image file

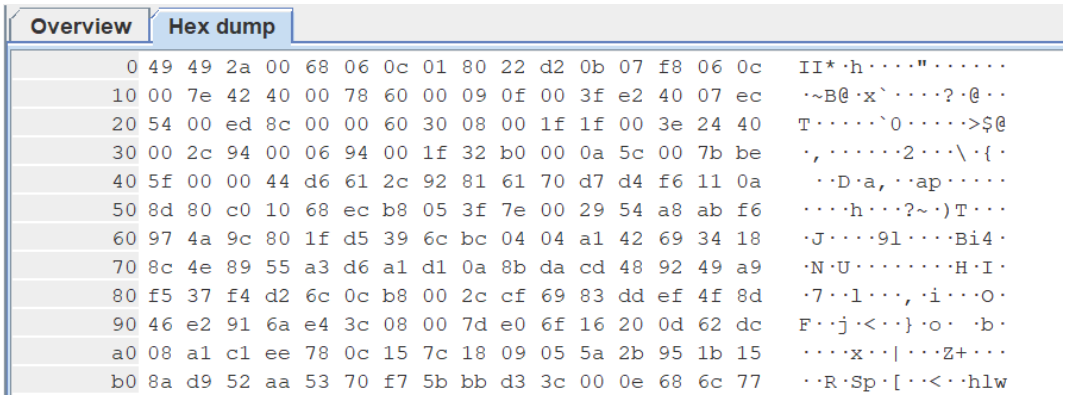


Figure 35 TIFF image file

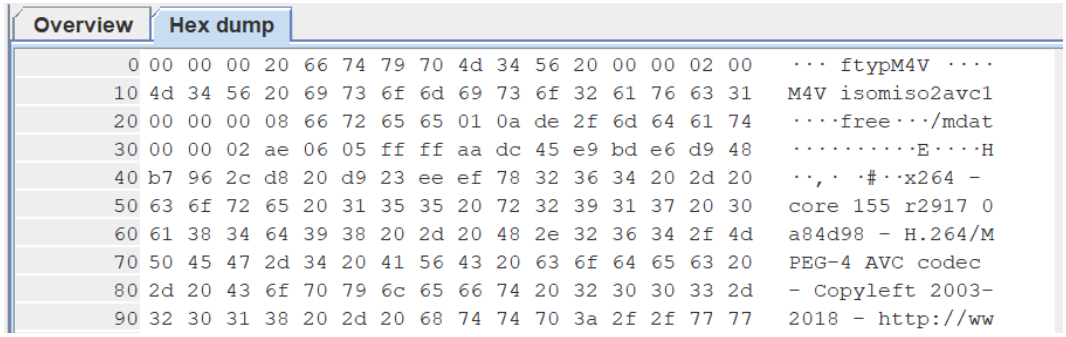


Figure 36 MP4 video with audio file

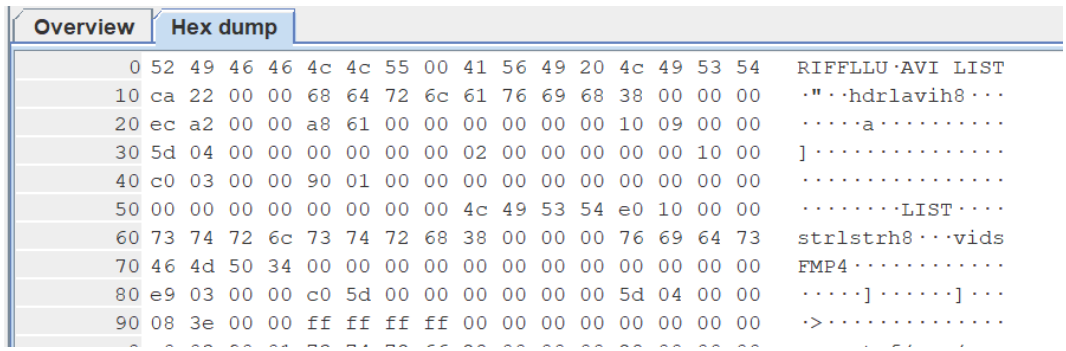


Figure 37 AVI video with audio file

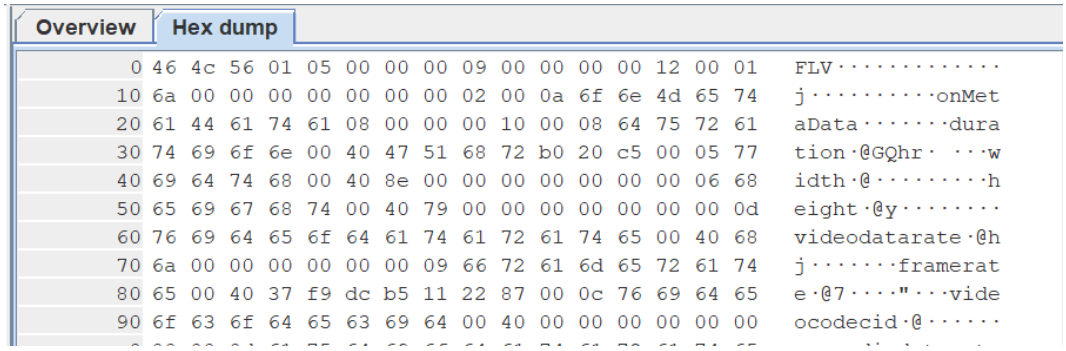


Figure 38 FLV video with audio file

It is left as an exercise for the reader to look in more detail at these types of files.

4.7 Scientific data

There are many diverse types of scientific data[71], ranging from simple tables of text to very complex structures. However, in many cases even complex data files can be described in a way which is much easier than for structures such as Word (.doc) files.

For those who do not work with data files, and even for those who do, the following examples may be interesting. If not then please at least look at the structures shown in the figures.

The Global Ozone Monitoring Experiment[72] (GOME) instrument was an atmospheric chemistry sensor onboard the second European Remote Sensing satellite (ERS-2, 1995 - 2011). ("GOME - European Space Agency")

The images created from the data included examples such as:

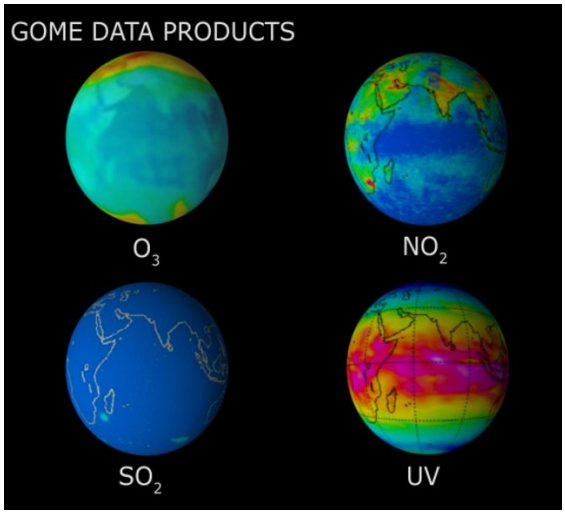


Figure 39 Examples images of GOME-2 data products

The data from GOME is archived by the European Space Agency[73] much of it in SAFE[74] format (Standard Archive Format for Europe) which is an implementation of the CCSDS/ISO XML Formatted Data Unit (XFDU) Structure and Construction Rules standard[75]. XFDU was written by the working group that created OAIS and is an extensible XML schema which has a place for all the elements required for an OAIS Archival Information Package.

As will be seen in Figure 40 the SAFE file is a container file - in this case a zipped folder which contains the additional schema elements specific to this type of data. You will see in

[71] See for example http://justsolve.archiveteam.org/index.php/Scientific_Data_formats which has a collection of scientific data formats and links to medical and other formats.

[72] See https://earth.esa.int/eogateway/instruments/gome

[73] https://earth.esa.int/eogateway/search?text=&category=Data&filter=GOME&subFilter=Data%20Description

[74] See https://earth.esa.int/eogateway/activities/safe-the-standard-archive-format-for-europe

[75] Available from https://public.ccsds.org/Pubs/661x0b1.pdf

the right-hand panel a reference to an XML namespace http://www.gael.fr/2004/12/drb/sdf, which indicates that the data is described using a data description language DRB, which is described in more detail in later sections, includes an XML schema annotation which allows its interpreter to extract bits from a data file to create the associated data elements, whether Strings, Integers, Reals etc.

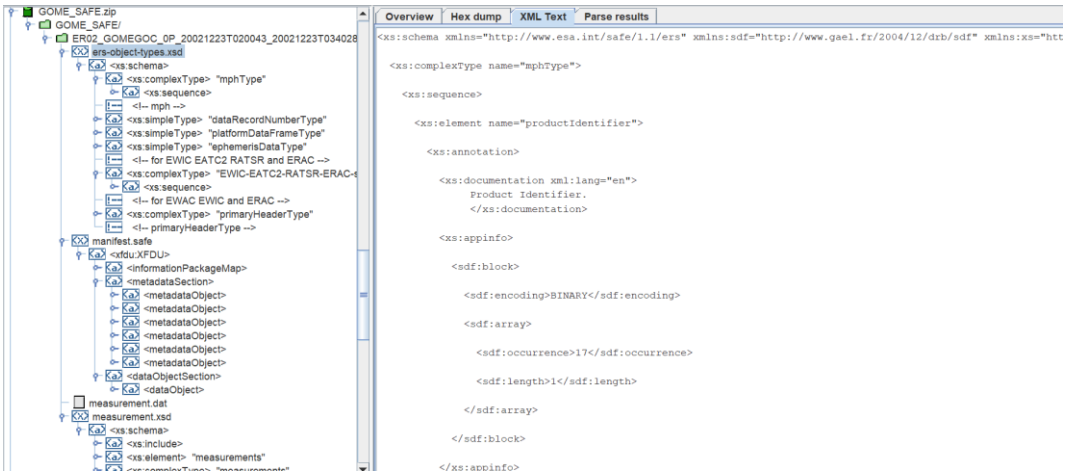


Figure 40 GOME SAFE file showing the data product specific schema elements

The next major element on the left-hand panel is the "manifest.safe" within which is the "XFDU schema, followed by the "measurement.dat," which is the actual data. In the right-hand panel in Figure 41 we can recognise a date and time "23-DEC-2022 02:00:43.974" but there are many non-printing characters, which presumably contain numbers and other data.

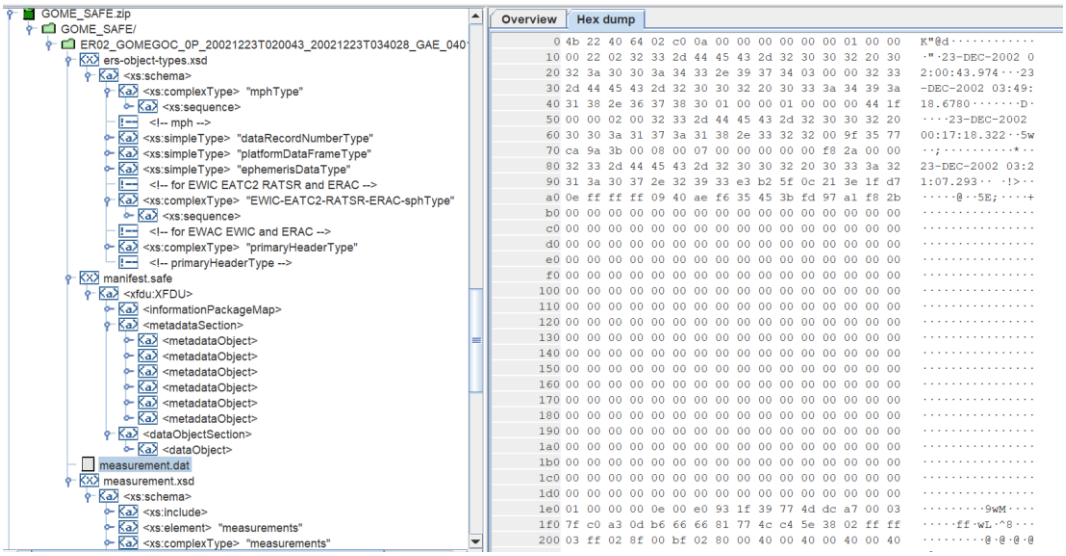


Figure 41 View inside a GOME SAFE measurement.dat file

The DRB interpreter allows one to extract data, which as pixel data for images, which can be used to construct complete images such as Figure 39.

4.7.1 FITS Table

One more, different, example of a table is an ASCII FITS table:

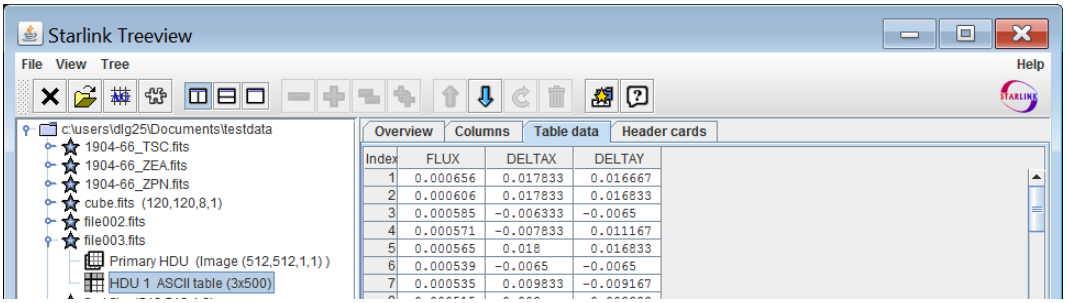


Table 14 Example FITS table

One can see the values in the table, with column names FLUX, DELTAX and DELTAY. What do they mean?

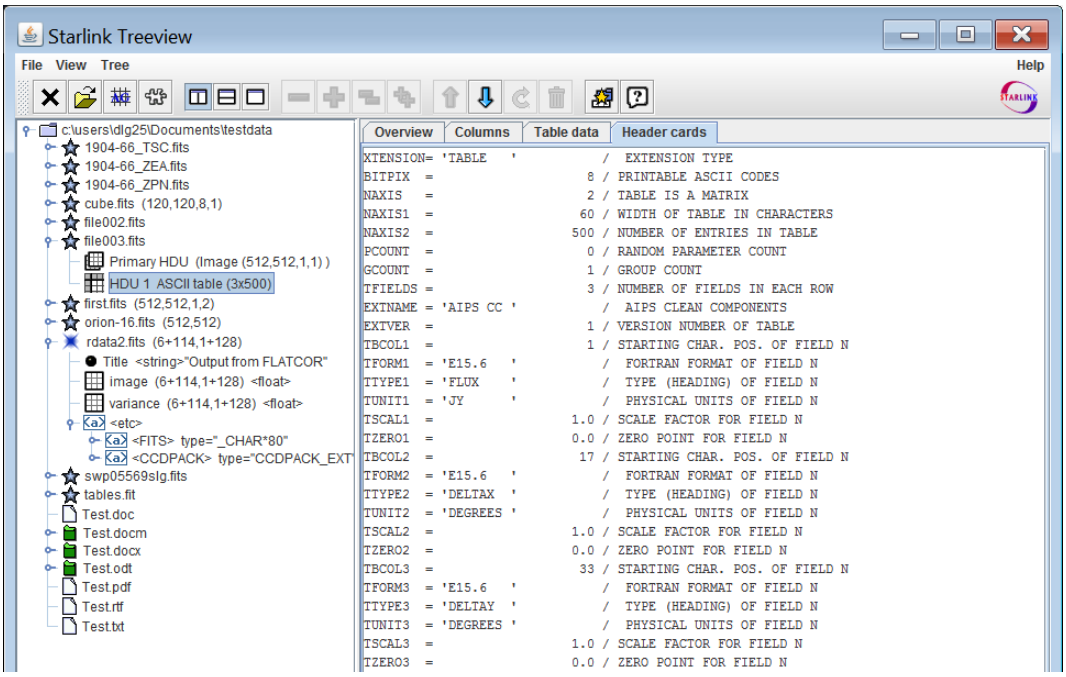


Table 15 Example FITS header

We see that the units of FLUX are Jy – which itself requires separate semantics to explain what this means. A collection of dictionaries used in FITS files from various organisations are available[76].

Looking down the headers of the associated image in that FITS file we can see that a great deal of extra information, including location in the sky and also information about processing history

[76] FITS keyword dictionaries see https://fits.gsfc.nasa.gov/fits_dictionary.html

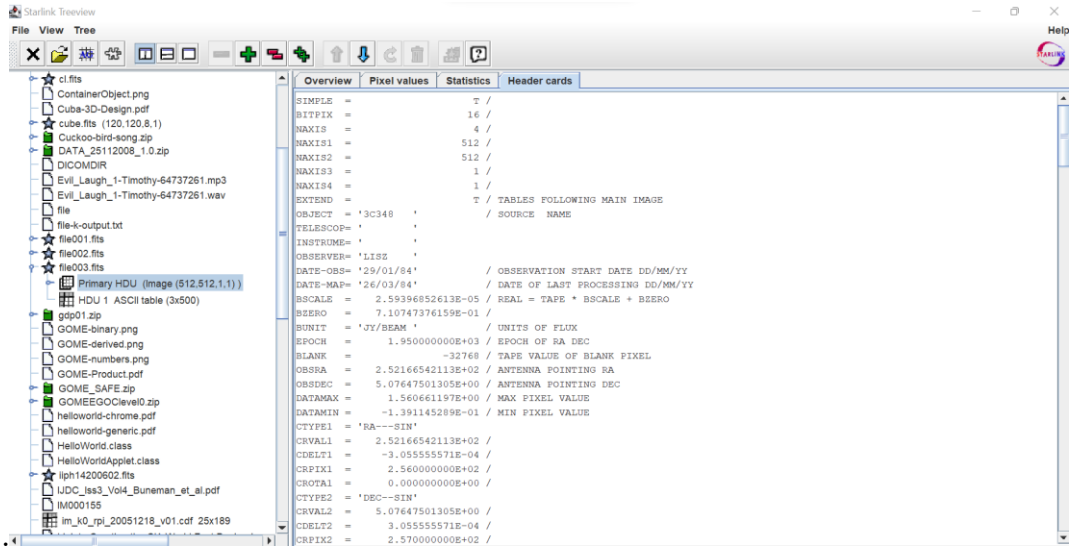


Figure 42 Header of FITS image which accompanies the table

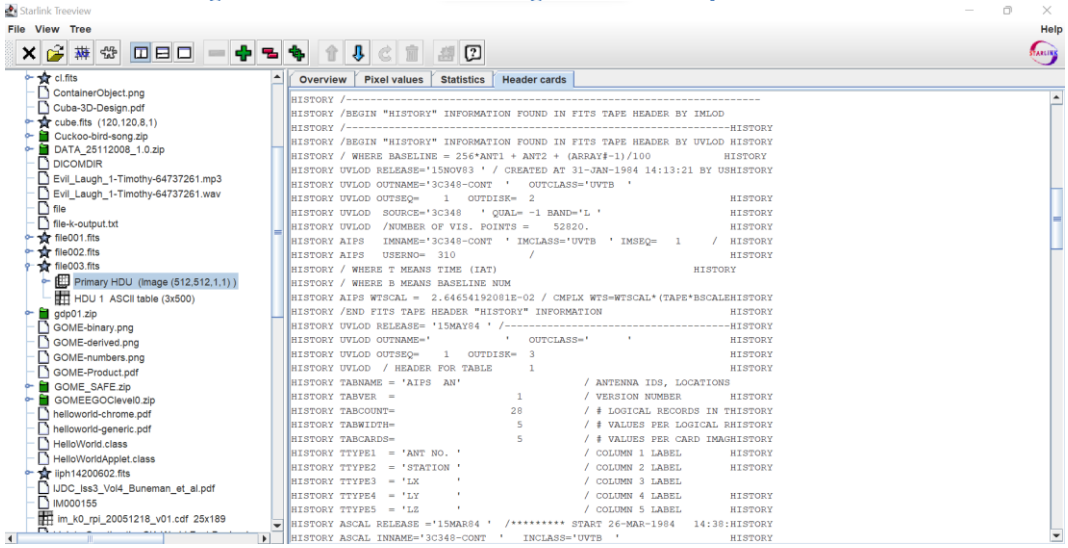


Figure 43 Header of FITS image showing processing history

The rather straightforward way that FITS files are constructed and described mean that it would be relatively easy to write software in future to extract the information from a FITS file. This is one of the reasons that the Vatican chose the FITS image format to archive many of its scanned images[77], and has provided information including semantic, of the specific keywords used [78], [79].

4.7.2 Formats for Gridded data

There are many specialised scientific data formats, and a number that are used quite generally across disciplines. Some have the word “common” or “universal” in their names, although one should not believe that claimed commonality and universality is true.

[77] Geerling, J., 2010, Vatican Secret Archive is Digitizing to Open FITS Format, https://www.opensourcecatholic.com/2010/vatican-secret-archive-is-digitizing-to-open-fits-format

[78] FITS files of the Library, https://www.vaticanlibrary.va/en/in-digitalizzation/fits-files.html

[79] See https://www.vaticanlibrary.va/en/the-collections/faq.html

4.7.2.1 BUFR, GRIB, CDF and NetCDF

Binary Universal Form for the Representation of meteorological data (BUFR[80]) and General Regularly distributed Information in Binary form (GRIB[81])are used in meteorology. ("BUFR format in a nutshell - confluence.ecmwf.int") Common Data Format (CDF[82]) has been used in space science data, but at various times special conventions have been used which impose specialised semantics on the data format.

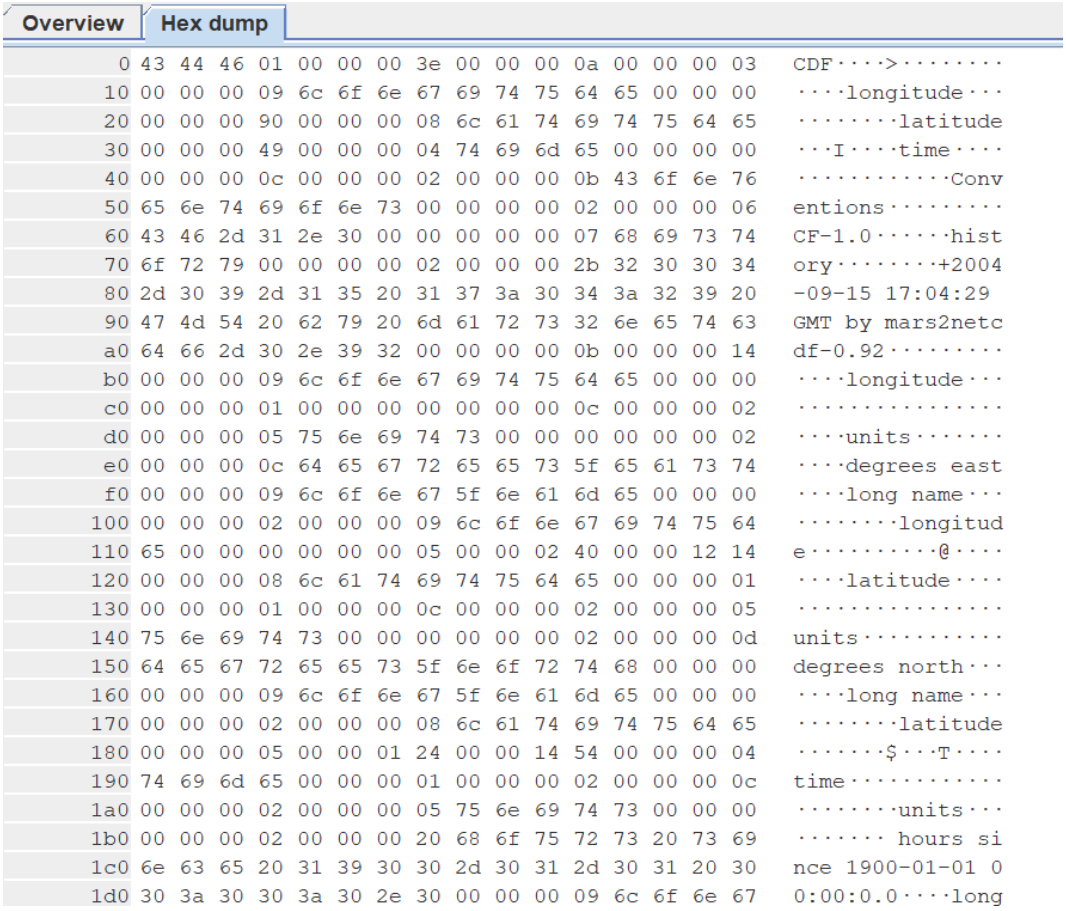


Figure 44 Example NetCDF file

An example of NetCDF[83] is shown in Figure 44, which gives an indication of the former relationship of NetCDF to CDF.

[80] BUFR see https://confluence.ecmwf.int/download/attachments/35752427/bufr_user_guide.pdf

[81] What are GRIB files and how can I read them, see https://confluence.ecmwf.int/display/CKB/What+are+GRIB+files+and+how+can+I+read+them

[82] Common Data Format – see https://cdf.gsfc.nasa.gov/

[83] NetCDF (Network Common Data Form) website https://www.unidata.ucar.edu/software/netcdf/

4.7.3 Formats for Hierarchical Data structures

4.7.3.1 HDF

Overview | Hex dump

```
0 89 48 44 46 0d 0a 1a 0a 00 00 00 00 00 08 08 00   ·HDF·····
10 04 00 10 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
20 ff ff ff ff ff ff ff ff 1b ef bc 07 00 00 00 00  ·····
30 ff ff ff ff ff ff ff ff 00 00 00 00 00 00 00 00  ·····
40 60 00 00 00 00 00 00 00 01 00 00 00 00 00 00 00  `·····
50 88 00 00 00 00 00 00 00 a8 02 00 00 00 00 00 00  ·····
60 01 00 b0 01 01 00 00 00 18 00 00 00 00 00 00 00  ·····
70 10 00 10 00 00 00 00 00 03 71 bc 07 00 00 00 00  ·····q·····
80 78 75 00 00 00 00 00 00 54 52 45 45 00 00 01 00  xu·····TREE····
90 ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff  ·····
a0 00 00 00 00 00 00 00 00 30 04 00 00 00 00 00 00  ·····0·····
b0 18 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
c0 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
d0 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
e0 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
f0 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
100 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ·····
```

Figure 45 Example HDF5 file

HDF5[84] is the latest incarnation of the HDF family. It supports n-dimensional datasets and each element in the dataset may itself be a complex object. ("Neurophysiological recordings from parietal areas of macaque brain ...") As will be seen below, it is widely used. HDF5[85] provides what looks like a directory structure within a single HDF5 object, to support which it defines a number of Virtual File Layers (VFLs) as illustrated in Figure 46.

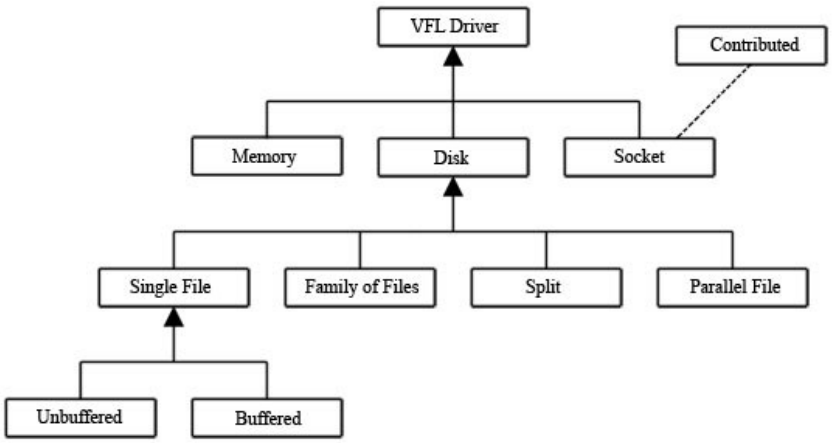


Figure 46 Conceptual hierarchy of VFL drivers

In addition, it supports a wide variety of datatypes, as shown in Figure 47.

[84] HDF5 website https://www.hdfgroup.org/solutions/hdf5/

[85] HDF5 data model and file structure, see http://davis.lbl.gov/Manuals/HDF5-1.8.7/UG/03_DataModel.html

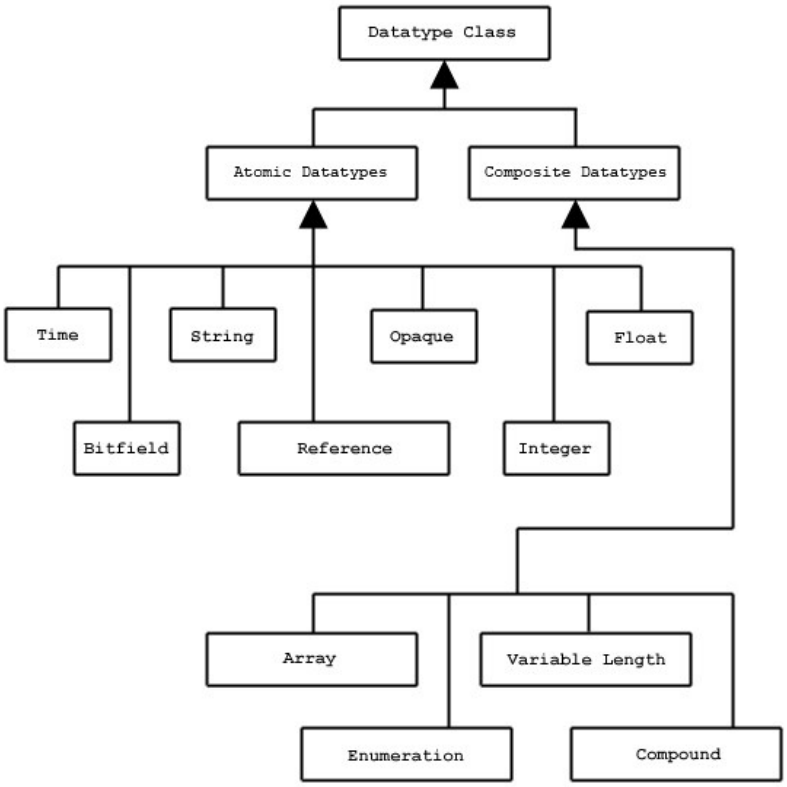


Figure 47 HDF5 Datatype classifications

4.7.3.2 HDS, NDF, HDX, NDX

The Starlink[86] astronomical software project developed complex, extremely flexible, hierarchical data formats[87]. The author has first-hand experience managing the Starlink project and is able to provide use useful insights. The first format, HDS[88] had key features which included:

- provision of a hierarchical organization of arbitrary structures, including the ability to store arrays of structures.
- the hierarchy is self-describing and can be queried.
- it gives the data author the ability to associate structures with an arbitrary data type.
- users can delete, copy, or rename structures within a file.
- Support for automatic byte swapping whilst using the native machine byte order for newly created output files

The original implementation has been replaced by an implementation on top of HDF5.

---

[86] The Starlink Project was a long running UK Project supporting astronomical data processing. It was shut down in 2005, but the software continued to be developed at the Joint Astronomy Centre until March 2015 and is now maintained by the East Asian Observatory. The code is open source. See lhttps://starlink.eao.hawaii.edu/starlink

[87] Jenness, T., "Learning from 25 years of the extensible N-Dimensional Data Format", Astronomy and Computing, vol. 12, pp. 146–161, 2015. doi:10.1016/j.ascom.2014.11.001. Available from https://arxiv.org/pdf/1410.7513.pdf

[88] Documentation for the Hierarchical Data System is available at https://starlink.eao.hawaii.edu/docs/sun92.htx/sun92se1.html

The flexibility of structure means that the analysis software must be tailored to the particular structure that has been created. It was realised that one must strike a balance between being very proscriptive in what applications can write out, e.g., simple FITS files, on the one hand, and allowing anarchy on the other. The most frequent problem was for applications to not understand the relationships between components, leading to erroneous processing, or for pieces of metadata not being understood or not correctly passed on to downstream applications.

The Starlink experience is that one needs some fairly simple rules which applications must obey, and that there should be some pre-defined components within which to hide additional structures in order to allow common operations to be dealt with uniformly and correctly. To simplify the use of the data, a logical structure with associated semantics was imposed on top of HDS, in the form of the N-Dimensional Data Format (NDF[89]).

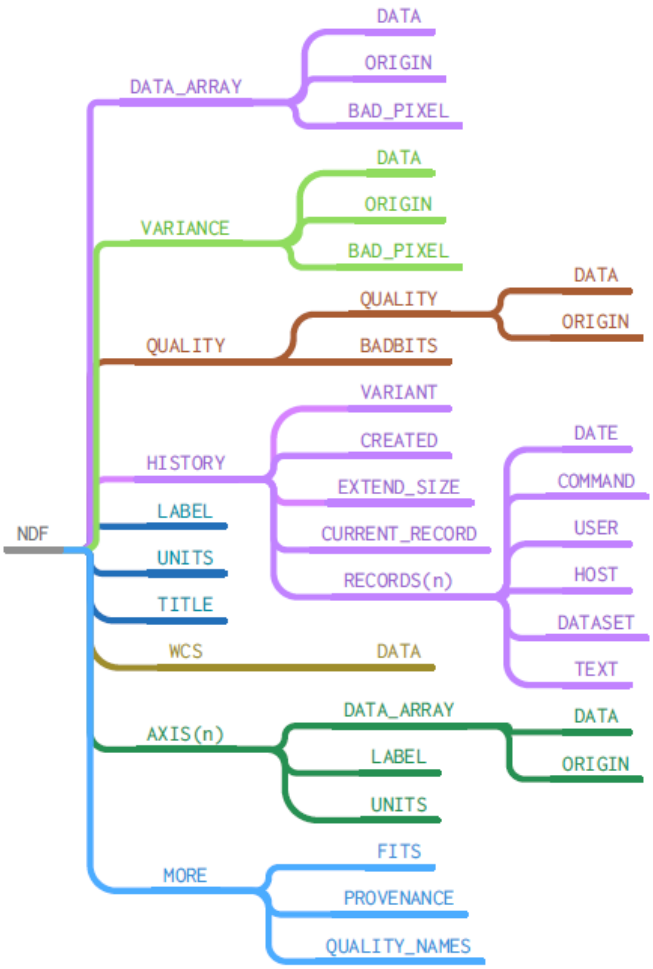


Figure 48 Schematic of the NDF hierarchy. All components are optional except DATA_ARRAY

As shown in Figure 48, beside the data itself, for example an image, one can include BAD_PIXEL values and whole images for VARIANCE, i.e. estimates of uncertainties or error bars, and QUALITY, i.e. indications of issues affecting the quality, providing those

[89] Extensible N-Dimensional Data Format (NDF) documentation https://starlink.eao.hawaii.edu/docs/sun33.htx/sun33.html

values for each pixel in the image. Processing the DATA_ARRAY image could then be accompanied by processing of the VARIANCE and QUALITY images.

It was later realised that the format should be open to the new tools and standards which are bound to be produced in the future outside the astronomical domain. It should also facilitate interoperation of applications—something which will become increasingly difficult as complex structures are generated. In addition, it must be possible for an application to adequately check the validity of a hierarchical file with which it is presented. The result was the creation of XML based formats HDX[90], [91] and NDX[92], the latter based on NDF. HDX may be regarded as essentially a manifest which points to a number of other objects in various formats, including FITS or other manifests.

4.7.4 Medical information

Much medical data is essentially scientific data in that it is analysed, for example for medical diagnoses.

There are many formats (with associated semantics) used in medical information. Common formats include Nifti[93], Minc[94] and DICOM[95]. Minc was initially based on the scientific formats NetCDF-(hence the name acronym MINC for "Medical Imaging NetCDF.") and then on HDF5. A view inside an example DICOM file (my own data) is shown in Figure 49.

[90] Giaretta, D., et al, "HDX Data Model: FITS, NDF and XML Implementation", 2003, Astronomical Data Analysis Software and Systems XII, ASP Conference Series, Vol. 295, 2003, available from http://www.star.bris.ac.uk/~mbt/papers/adassXII-O7-4.pdf

[91] HDX Overview, https://www.astro.gla.ac.uk/users/norman/note/2003/hdx-overview/

[92] Giaretta, D., et al., "Starlink Software", 2004, Astronomical Data Analysis Software and Systems XIII, ASP Conference Series, Vol. 314, 2004, available from https://articles.adsabs.harvard.edu/full/seri/ASPC./0314//0000835.000.html

[93] Originating in the Neuroimaging Informatics Technology Initiative see https://radiopaedia.org/articles/nifti-file-format?lang=gb

[94] Used in neuroimaging, see Vincent RD, Neelin P, Khalili-Mahani N, Janke AL, Fonov VS, Robbins SM, Baghdadi L, Lerch J, Sled JG, Adalat R, MacDonald D, Zijdenbos AP, Collins DL and Evans AC (2016) MINC 2.0: A Flexible Format for Multi-Modal Images. *Front. Neuroinform*. 10:35. doi: 10.3389/fninf.2016.00035 http://doi.org/10.3389/fninf.2016.00035

[95] The DICOM standard – see https://www.dicomstandard.org/

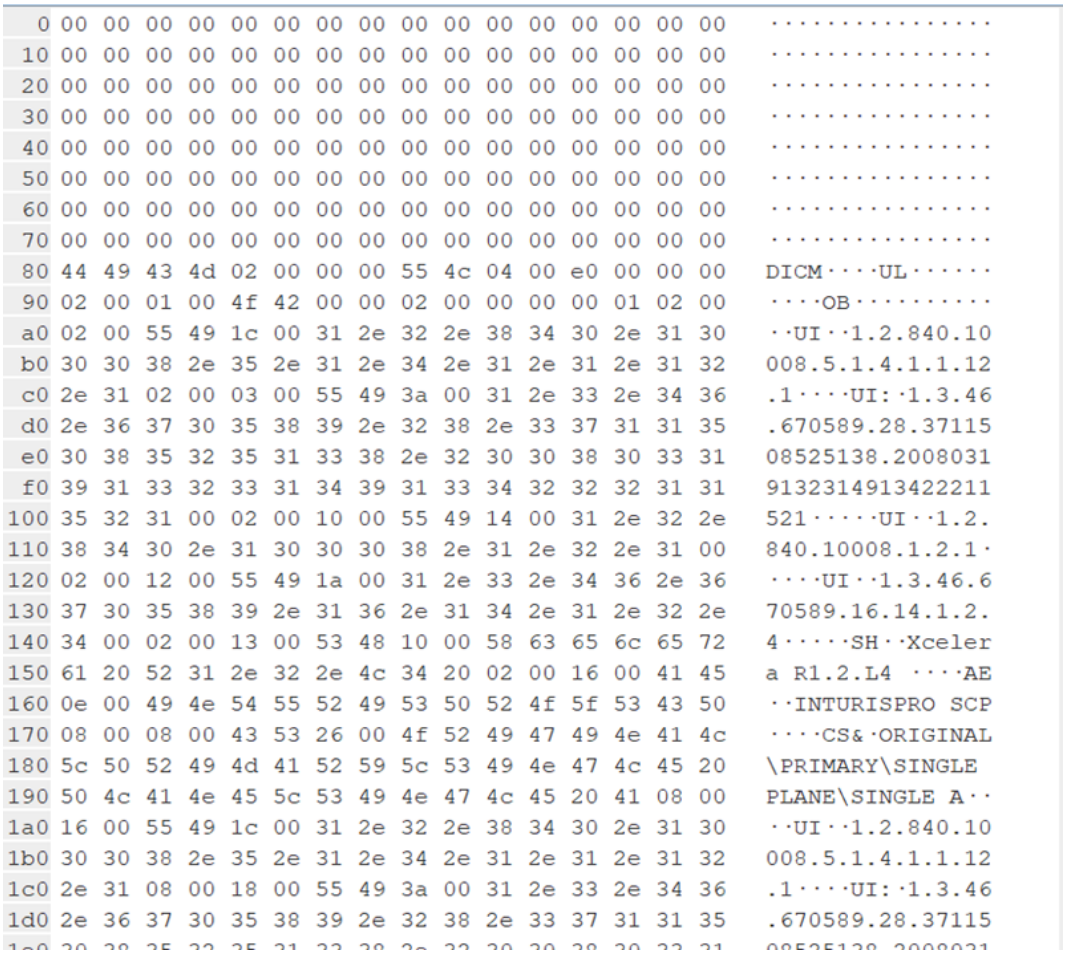


Figure 49 Example DICOM file

Enough for now about the bits and bytes. We can turn to the bigger picture in the next chapter.

4.8 Limitations of this chapter

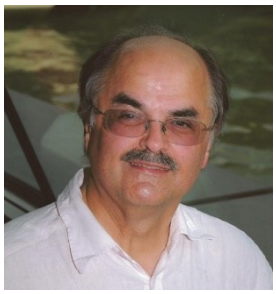
One can only look at a limited number of examples, and one must trust that we will not be surprised with strange objects.

We cannot explain every detail of the internal structure of complex objects such as a Word file in this chapter. On the other hand, a full description must have been available to whoever wrote the MS Word application, although whether it is not clear whether that description is available to anyone else.

What is important is that the reader should be aware of some the details of files which are normally hidden or ignored because it is possible that these details may come back to bite us some day, whether in terms of network resources or effects of transformations to different formats.

About the Author

David Giaretta has worked in digital preservation since 1990 and has led many of the most important developments in this area.



He chaired the panel which produced the OAIS Reference Model (ISO 14721), the "de facto" standard for building digital archives, and made fundamental contributions to that standard. He led the 2012 update of OAIS and the current update.

He leads the group which produced the ISO standard for audit and certification of trustworthy digital repositories (ISO 16363), and IS016919, which are fundamental to the setting up the certification process; details are available on the PTAB website (www.iso16363.org).

He has led a number of large digital preservation projects, representing an investment by the EU and more than 50 partner organisations of several tens of millions of Euros. These projects include CASPAR, PARSE.Insight, APARSEN and SCIDIP-ES – details available here[96]. These build on his experience working in and leading large data digital repositories and software systems.
Involved with the Alliance for Permanent Access[97] (APA) from its start to its establishment, David Giaretta became the Director of the APA in July 2010.
In 2011 his book "Advanced Digital Preservation" was published; a full list of publications are available here[98].

Most organisations depend on their digitally encoded intellectual capital. Backup plans can help ensure business continuity in case of disaster but how can your organisation ensure that it benefits from its digital capital?

[96] http://giaretta.org/digital-preservation/projects/

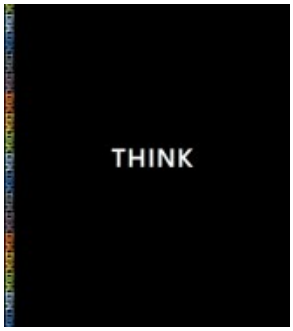
[97] http://www.alliancepermanentaccess.org/

[98] http://giaretta.org/publications/

Award from President Reagan to the International Ultraviolet Explorer Team

Award for work on Hubble Space Telescope

In 2003 he was awarded an MBE for services to Space Science.



Featured in IBM Think 2006

September 2012 David was awarded the Emmett Leahy Award for Outstanding Contributions to the Information and Records.

August 2013 David Giaretta was appointed special advisor for digital preservation to Renmin University, Beijing

Background

He gained his BA, MA, MSc, and a doctorate in Theoretical Physics from Oxford University.

He chairs the CCSDS/ISO panel which wrote the OAIS standard (ISO 14721:2002) and led the updates in 2012 and 2025. He then worked on many of the follow-on standards listed in the OAIS roadmap. Work on certification was carried forward by the Task Force on Digital Repository Certification, of which he was a member. This group produced, in 2007, the TRAC document which was the initial draft on certification. As planned, this draft was taken back into CCSDS/ISO to a panel which he chaired, which refined and reorganised it to make it more suitable for audits. In 2012 this was published as ISO 16363, which completely superseded TRAC. The panel then produced ISO 16919:2014 which plays the vital role of specifying the requirements on bodies which provide audit and certification, fitting into the international ISO process of audit and certification. He led the updates of ISO 16363 and ISO 16919 in 2025.

He has worked on a number of astronomical satellites, being UK resident astronomer for

IUE[99] and the calibration lead in the USA on the Faint Object Camera[100] of the Hubble Space Telescope[101] and has extensive experience in planning, developing and running scientific archives and providing and managing a variety of services to large numbers of users.
Dr Giaretta has published many scientific papers in refereed journals and given presentations at many international conferences, scientific as well as technical. In addition, he has broad experience in e-Science. He is a member of the programme committee of the PV[102] series of conferences which focuses on preservation and adding value to scientific and technical data and several other conference series.
In 2005 he co-organised the Warwick workshop[103] "Digital Curation and Preservation: Defining the research agenda for the next decade" which has proved to be very influential in the UK and internationally. More recently he was rapporteur for the EU High Level Group on Scientific Data whose report Riding the Wave[104] put forward a vision for 2030.
Dr Giaretta was a founding Associate Director for Development of the UK Digital Curation Centre[105]. He then led a number of EU projects, closely linked to the APA, focussed on digital preservation. These included CASPAR which implemented much of the first APA research plan, addressing fundamental issues of digital preservation, producing prototype infrastructure components and tools for preservation of all types of digitally encoded information together with evidence of their effectiveness. PARSE.Insight identified users' views on the main threats to, and their attitudes towards, preservation.
APARSEN brought together commercial and public organisations across Europe to come to a common vision for research in digital preservation. The SCIDIP-ES project assessed key infrastructure components, based on those prototyped by CASPAR, to help organisations be more trustworthy as preservers of digitally encoded information.

He founded and co-chairs the Research Data Alliance[106] group on Active Data Management Plans and the group on Preservation e-Infrastructure.

He, together with colleagues, explored Information Preservation Sustainability in a 2024 UNESCO report[107].

[99] For details see https://science.nasa.gov/missions/iue and https://www.esa.int/Science_Exploration/Space_Science/IUE_overview

[100] See https://www.stsci.edu/hst/instrumentation/legacy

[101] See https://www.nasa.gov/mission_pages/hubble/main/index.html

[102] List of PV conferences see http://www.alliancepermanentaccess.org/index.php/community/conferences/pv-conferences/

[103] For information see https://www.dcc.ac.uk/events/workshops/digital-curation-and-preservation-defining-research-agenda-next-decade

[104] Riding the Wave. How Europe can gain from the rising tide of scientific data https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data

[105] https://www.dcc.ac.uk/

[106] See https://www.rd-alliance.org/

[107] Giaretta, David, Lowry, James, & Kenmoe, Michel. (2024). Indicators of Information Preservation Sustainability. UNESCO and University of Ghana Press. https://unesdoc.unesco.org/ark:/48223/pf0000392169