Associated Knowledge Preservation Best Practices

CEOS

Data Stewardship Interest Group

Doc. Ref.: CEOS/WGISS/ DSIG/AKPBP

Date: January 2017

Issue: Version 1.0

Document Status Sheet

Issue	Date	Comments	Editor
0.1	January 2016	First draft of Table of Content issue	Iolanda Maggio
0.2	April 2016	First draft of the Best Practices	Iolanda Maggio
0.3	September 2016	New RIDs from LTDP WG members implemented	Iolanda Maggio
1.0	January 2017	Final review of the document	Iolanda Maggio

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Purpose of the document

This document aims to provide recommendations and best practices for the preservation of Earth Observation (EO) Associated Knowledge.

1.2 Intended Audience

This document is intended to assist data holders in Earth Observation (EO) data centres in the task of ensuring Earth Observation space Associated Knowledge long-term preservation.

1.3 How to use these Best Practices

In accordance with the CEOS Best Practices, the Associated Knowledge preservation needs to be tailored for the specific mission relevancy and requirements, taking into consideration the user community needs, cost / benefit analysis, preservation objectives, risk assessment, Intellectual Propriety Right (IPR), and SW dependencies.

1.4 Document Overview

This document is divided into:

Section 1: Introduction

Section 2: Background

Section 3: Objectives and needs

Section 4: Associated knowledge elements and formats

Section5: Recommendations

Annex A: Use cases Scenarios

Annex B: Software Preservation Techniques

1.5 Acronyms

Acronym	Description
CEOS	Committee on Earth Observation Satellites
FITS	Flexible Image Transport System
IPR	Intellectual Propriety Right
MS	Microsoft
PDF/A	Portable Document Format
PNG	Portable Network Graphics
WGISS	Working Group on Information Systems and Services

1.6 Related Documents

1.6.1 Applicable Documents

Applicable Document ID	Document Title	Reference
AD-1	CEOS, "Long Term Preservation of Earth Observation Space Data: Preservation Workflow"	CEOS/WGISS/DSIG/PW
AD-2	CEOS, "EO Preserved DataSet Content"	CEOS/WGISS/DSIG/EOPDSC
AD-3	POCOS, "The Preservation of Complex Objects Volume 1: Visualisations and Simulations"	ISBN 978-1-86137-6305

1.6.2 Reference Documents

The following documents, though not formally part of this document, amplify or clarify its content.

Reference Document ID	Document Title	Reference	Availability
	PRESERVING VIRTUAL WORLDS FINAL REPORT		
	Digital preservation and curation - the danger of overlooking software		

2 BACKGROUND

Digital Preservation represents the management and maintenance of *digital objects* so they can be accessed and used by future users.

In the context of Earth Observation, digital objects are composed of**:**

- ✓ **Data Records**: these include raw data and/or Level-0 data, higher-level products, browse images, auxiliary and ancillary data, calibration and validation data sets, and descriptive metadata;
- ✓ **Associated Knowledge**: this includes all the *Tools* used in the Data Records generation, quality control, visualization and valorisation, and all the *Information* needed to make the Data Records understandable and usable by the Designated Community.

Long-term accessibility and exploitability of Earth Science data requires that not only sensed data, but also associated knowledge, needs to be properly preserved and made accessible.

Information technology is changing rapidly and this change also affects digital data from Earth Observation missions. Risks include the corruption of the bit-stream, obsolescence of the file format, extant hardware and operating environments that make data unreadable on the physical and logical level. On the other hand, insufficient documentation regarding the data, the inability to discover the data, or service compatibility can also prevent their re-use.

Digital objects need a hardware and software environment in order to be managed. These environments can be complex and in some cases, distributed. The technological evolution makes them obsolete in a short span of time so, ways of preserving both media and their execution context must be found. For some digital objects, such as software programs, the absence of source code may be a problem. Legal aspects, such as copyrights or copy protection mechanisms, can make this even more complex.

3 OBJECTIVES AND NEEDS

Digital Preservation consists of the management and maintenance of digital objects so they can be accessed and used by future users. It is important to start thinking about digital preservation early in the life cycle of a digital object because they have significantly shorter life spans. Therefore, by considering the preservation of a digital object early on, even when it is created, a great deal of time and stress is saved later on, when trying to retrieve the information an object holds. In this sense, digital preservation, and especially early digital preservation, is important not only for personal data management, but also for large repositories that manage a lot of digital objects.

Digital Preservation is frequently focused on long term use, which can be quite difficult to achieve considering how fragile digital objects can be.

The main purpose of this document is to help data owners in preserving information and tools through recommendations and guidelines.

In accordance with the CEOS Best Practices, the Associated Knowledge preservation needs to be tailored for the specific mission relevancy and requirements, taking into consideration the user community needs, cost / benefit analysis, preservation objectives, Intellectual Propriety Right (IPR), and SW dependencies.

The data manager should tailor the Associated Knowledge preservation to meet the needs of the specific mission, stating which knowledge should be preserved during each phase of the Preservation Workflow in accordance with [AD-1] and maintain the Preserved Data Set Content inventory table with information, and software available under configuration, in accordance with [AD-2].

The Associated Knowledge preservation, through the Preservation Workflow implementation can be performed to historic, current, and future Earth observation missions. For historic missions, difficulties may arise in recovering all the relevant information and tools. For current missions, preservation activities should be initiated while the mission is still in operation in order to recuperate all relevant information. For future missions the definition of long-term preservation strategies and implementation aspects should ideally be planned for or initiated during the mission preparation phases. This will facilitate the availability and usability of data records and associated knowledge during the mission lifetime and in the long-term, and will reduce associated consolidation and preservation costs.

4 ASSOCIATED KNOWLEDGE ELEMENTS AND IDENTIFIED FORMATS

A non-exhaustive list of Associated Knowledge types is presented below:

- ✓ Information:
  - Documentation
  - Images
  - Metadata files (information on creation, access rights, restrictions, preservation history, and rights management)
  - Multimedia (Video/Audio)
  - Workflows
  - Bi directional links
  - Schemas
  - Emails
- ✓ Software/Tools:
  - Software Applications:
    - Data Product generation
    - Quality control
    - Product visualization
    - Value adding
  - SW related "IT Infrastructure":
    - Compiler
    - Programming language
    - Storage system
    - Operative System
    - Libraries
    - Databases

4.1 Information

Some identified features of Information formats for Digital Preservation are presented below:

- Freely available.
- Limitation in patents or licenses on the format.
- Ubiquitous and in wide use.
- Have an extensive feature set.
- Endorsed by other established repositories.

- Variety of writing and rendering tools available for the format.
- Interoperable, with tools that allow the conversion to other formats.
- Inclusion of error-correction capabilities.
- For image, video and audio information, lossless formats are desirable.
- Support a stable mechanism for metadata management.
- Data integrity features, based on flexible digital signatures for existing and future cryptographic methods.

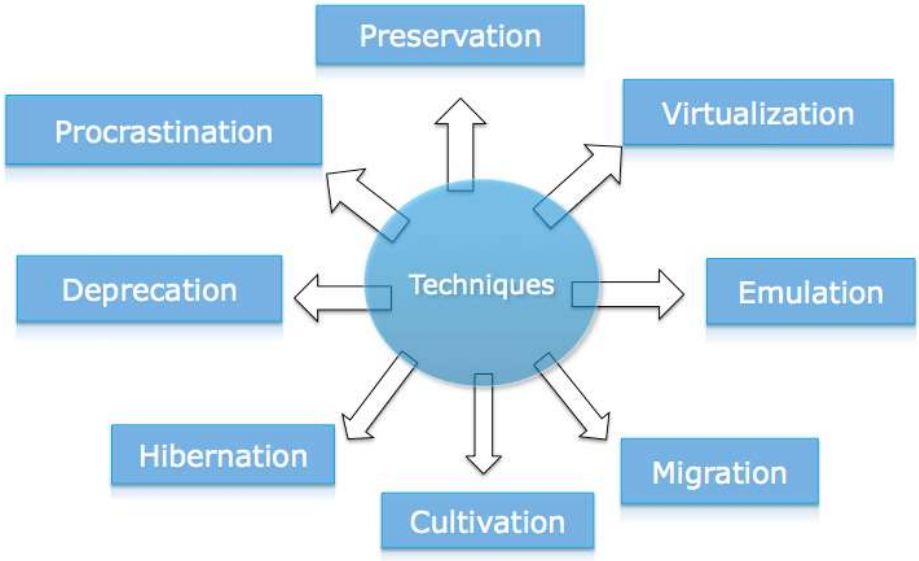
Self-describing capabilities are very desirable, especially for files that store big data sets.

The existing and possible preservation format of Information Preservation are listed below:

- ✓ Text documents (often MS Word, Excel Files, txt, etc.) can be preserved as:
  - PostScript, PDF/A, DSSSL, RTF, ASCII, SGML, TIFF, CGM, FITS
- ✓ Images (JPEG, TIFF, PNG, FITS, etc.) can be preserved as:
  - Loss of Quality JPEG, JPEG2000
  - Lossless compression TIFF, PBM, PNG, FITS.
- ✓ Metadata (ASCII, XML, SGML, etc.) can be preserved as:
  - ASCII, the most durable format for metadata because it is widespread, backwards compatible when used with Unicode (superset of ASCII), and utilizes human-readable characters, not numeric codes.
  - For higher functionality, SGML or XML should be used.
- ✓ Multimedia (AVI, QuickTime, MPEG, WMV, MJ2) can be preserved as:
  - MJ2 and MPEG-A.

4.2 Software/Tools Preservation

The approaches for Software preservation are presented in the figure below:

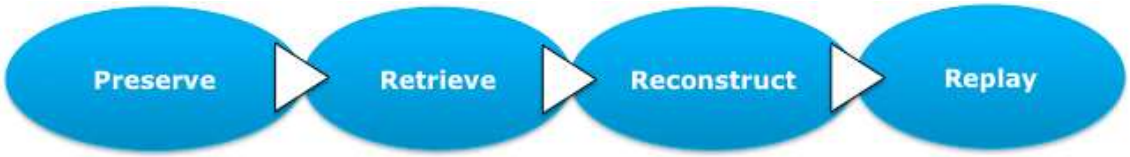


As already stated, during the initial dataset appraisal, a decision on the approach to be followed in order to handle and preserve the mission data, information and the related SW, is taken.

This decision depends on:

- Mission relevance, temporal and geographical coverage, size, storage media and archiving format, technical aspects and cost.
- Software Preservation approach can be based on different strategies for different missions (e.g. virtualization, periodic migrations, hibernation).

In accordance with [AD-3] the Software preservation steps are described in the diagram below.



Preserve: A copy of the software to be stored for long-term preservation. There should be a strategy to ensure that the storage is secure and maintains its authenticity over time, with appropriate strategies for storage replication, media refresh, format migration, etc.

Retrieve: In order to retrieve the preserved SW at a date in the future, it needs to be clearly labelled and identified, inventoried or catalogued. This should provide search on its function and origin (provenance information).

Reconstruct: The preserved product can be reinstalled or rebuilt within a sufficiently close environment to the original, so that it will execute satisfactorily. For software, this is a particularly complex operation, as there are a large number of contextual dependencies to the software execution environment which are required to be satisfied before the software will execute at all.

Replay: In order to be useful at a later date, software needs be replayed or executed, and performed in a manner which is sufficiently close in its behaviour to the original. As with reconstruction, there may be environmental factors which may influence whether the software delivers a satisfactory level of performance.

The Software preservation is a highly complex activity with a lot of factors which need to be considered, because:

- ✓ Software is not easy to be defined.
- ✓ Software has lots of different components and dependencies.
- ✓ Software operates in a complex environment.
- ✓ Software represents a very large topic.

When the software preservation activities start, various attributes should be analysed:

- **Functionality (**What it does and what data it depends on, etc.)
- **Environment (**Platform, operating system, programming language, versions, etc.)
- **Dependencies (**Compilation, Standard libraries, etc.)
- **Composite digital object (**Collection of modules, Specifications, configuration scripts, test suites, documentation)
- **Architecture (**Client/server, storage system, input/output)
- **User interaction (**Command line, User Interface, User model)

5 RECOMMENDATIONS

This document addresses the main "areas" in which the recommendations should be applied in order to guarantee the Associated Knowledge elements preservation over time.

Some Use Cases are provided in Annex A in order to describe the Earth Sciences and Earth Observation mission context.

5.1 Recommendations Organization

The Associated Knowledge preservation recommendations are divided in general and specific areas of recommendations. Three areas are considered:

- ✓ Future Missions;
- ✓ Historical Missions;
- ✓ Current Missions.

5.2 Recommendation Formatting

Each recommendation in this document is assigned a unique identifier.

The recommendation ID scheme follows the pattern:

REC_< SOURCE>_<AREAS>_<ELEMENT>xxx

where:

- **REC** is a constant value for all recommendations.
- **<SOURCE>**

SOURCE	Type
GEN	General
SPEC	Specific

- **<AREAS> if** the source is Specific, it denotes the areas of the mission status

AREAS	Type
FUT	Future
HIS	Historical
CUR	Current

- **<ELEMENT>**

ELEMENT	Type
ALL	Associated Knowledge
INF	Information
SW	Software

- **xxx** Sequential Number

5.3 General Recommendations

[REC_GEN_ALL_01]

In order to be able keep track of an organisation's digital assets related to Earth Sciences and Earth Observation missions, it is recommended to collect and store the metadata of the digital objects in use, or deemed to be preserved, according to international metadata standards and practices for preservation (e.g. PREMIS, Dublin Core, etc.).

[REC_GEN_INF_01]

In order to guarantee a reliable preservation of information, several formats are identified. When a digital object is identified as mandatory, two different formats should be used.

- Mission documentation (***mandatory***): the preservation formats considered should be PDF/A and FITS.
- SW and HW documentation (***recommended***): the preservation formats to be considered are PDF/A and/or FITS.
- Images (***mandatory***) related to EO missions (e.g. missions quick-looks, paper images etc.): the preservation format considered should be TIFF and FITS. (FITS preferred for flexibility in metadata editing. It allows handling more complex metadata schemas).
- Metadata files (***mandatory***): the main selected preservation formats are XML and ASCII.
- Multimedia files (***mandatory***): the main recommended formats (e.g. Videos) are MJ2 and MPEG-A

[REC_GEN_SW_01]

Emulation, virtualization requires using other software in order to preserve the software of interest. Despite the approach (emulation or virtualization) chosen and implemented to preserve the mission SW and tools, it will have to be refreshed and upgraded periodically.

[REC_GEN_SW_02]

If Emulation is selected for SW preservation the emulator should have the following characteristics:

- The emulator is based upon freely available source code and appropriate licensing.
- The emulator is actively maintained.
- There is reasonable internal and external documentation.
- The emulator interface is easy to use by non-technical players.
- The emulator supports a wide range of performance and tuning options.
- The emulator is robust and provides a believable level of fidelity when compared to the original game experience.

[REC_GEN_SW_03]

If Virtualization is selected for SW preservation the emulator should have the following characteristics:

- The virtualization technology is based upon freely available source code and appropriate licensing. In cases where the source was not available, the virtualization technology was being developed by a company dedicated to virtualization.

- The virtualization technology is actively maintained.
- There is reasonable documentation for the project.
- The virtualization interface is easy to use.
- The virtualization technology is robust and provides a believable level of fidelity when executing legacy software.

[REC_GEN_SW_04]

If communities of scientists write their own specialist processing programs, such as for remote sensing discipline, the Cultivation approach can be applied in order to keep the software 'alive'. Cultivation can benefit from the use of Open Source Software.

5.4 Specific Recommendations

5.4.1 Future Missions

[REC_SPEC_FUT_ALL_01]

Recommendations about software engineering best practice such as clear licensing; clear documentation; use of commonly adopted and modern programming languages; modular design; revision management and change control; established software testing regime and validated results; separation between data and code; and clear understanding of dependencies, all facilitate software preservation. If the already existing Software Engineering Best Practices cover all preservation requirements, only a statement, with the standards' references, should be highlighted.

[REC_ SPEC_FUT_SW_01]

Open source software should be preferred for any development in future missions.

In order for future users to be able to use the software, it is critical that they have permission to access the code, to make any changes necessary to get it working, and they should to be able to do so without any sort of implied warranty from the original users (particularly as the rest of the software ecosystem will have moved on). Making software available as Free Software or Open Source Software is a simple way of ensuring that this is the case.

[REC_SPEC_FUT_SW_02]

If licensed software is used for new developments, the copyright must be clearly identified and publicized, in order to facilitate future use and preservation.

[REC_ SPEC_FUT_SW_03]

When a new SW is developed, a relation network should be created in order to trace the provenance. The SW needs to be bi-directionally linked to the data-records, documentations, sensors and missions.

[REC_ SPEC_FUT_SW_04]

As well as preserving the source code of the software, it is worth considering whether the history of changes to the software is also in scope for preservation. Most modern software is developed using a revision control system (E.g. Git, Subversion, others) that keeps track of every change made to the

software over time. This means it is possible to preserve not just the state of software at the time of preservation, but also its entire history up to that point. This may have future cultural heritage value, as it makes it possible to see the contributions of individuals, as well as the responses of the software developers to external events as expressed through changes to the code. It also makes it clearer when, and why, particular changes to the code were made.

5.4.2 Historical Missions

[REC_SPEC_HIS_ALL_01]

Prior to starting the preservation activities, several important, non-exhaustive, aspects should be taken into consideration: mission relevancy, preservation scenarios, amount of funding, cost/benefit analysis, designed community requirements, Intellectual Propriety Rights (IPRs), and SW dependencies. Following the assessment, if it is discovered that there are issues related to the recovery of Associated Knowledge, a justification should be openly declared.

[REC_SPEC_HIS_SW_01]

If there is no funding and no mission interest because data records are not unique or already covered by another organisation/owner, then the preservation, hibernation or the procrastination methodology, should be implemented (To be decided by the Data Holder).

[REC_SPEC_HIS_SW_02]

If a historical SW has been maintained by means of emulation, or virtualization, or migration etc., all information regarding this preservation activity should be collected in order to trace the implemented changes and evolutions. Revision Control System is recommended (e.g. Git or Subversion).

[REC_SPEC_HIS_SW_03]

If the Software to be preserved needs to maintain the original fidelity, the Virtualization or Emulation approaches should be implemented, without making any changes to the original software.

[REC_SPEC_HIS_SW_04]

If the software has an intrinsic value and it can be a valuable historical resource (e.g. if it is the first example of its type, or it was a fundamental part of a historically significant event), the software has inherent heritage value and should be preserved as it is (Preservation approach).

[REC_SPEC_HIS_SW_05]

If the software can't be separated from the data or digital object (For example, if the software and the data form an integrated model, where the data by itself is meaningless) it must be preserved together with the data.

[REC_SPEC_HIS_SW_06]

If the owners and the developers of the code are available, the source code and the relevant hardware are accessible and if there are no issues on licensing or rights management, then the migration of the software can be followed.

[REC_SPEC_HIS_SW_07]

If SW is solely used internally by an organization, where there is no use or interest from external users, and it must perform exactly as the original, then the preservation approach can be followed.

5.4.3 Current Missions

[REC_SPEC_CUR_ALL_01]

For Current Missions, a mixed approach can be followed, where recommendations for both Historical and Future missions can be pursued. For example, for new developments, Future Missions recommendations should be implemented; otherwise Historical recommendations can be followed.