D6.6.1 Current state of 3D object digital preservation and gap-analysis report

DURAARK

FP7 – ICT – Digital Preservation
Grant agreement No.: 600908

Date: 2014-01-22

Version 1.0

Document id. : duraark/2014/D.6.6.1/v1.0

Grant agreement number	:	600908
Project acronym	:	DURAARK
Project full title	:	Durable Architectural Knowledge
Project's website	:	www.duraark.eu
Partners	:	LUH – Gottfried Wilhelm Leibniz Universitaet Hannover (Coordinator) [DE] UBO – Rheinische Friedrich-Wilhelms-Universitaet Bonn [DE] FhA – Fraunhofer Austria Research GmbH [AT] TUE – Technische Universiteit Eindhoven [NL] CITA – Kunstakademiets Arkitektskole [DK] LTU – Lulea Tekniska Universitet [SE] Catenda – Catenda AS [NO]
Project instrument	:	EU FP7 Collaborative Project
Project thematic priority	:	Information and Communication Technologies (ICT) Digital Preservation
Project start date	:	2013-02-01
Project duration	:	36 months
Document number	:	duraark/2014/D.6.6.1
Title of document	:	D6.6.1 Current state of 3D object digital preservation and gap-analysis report
Deliverable type	:	Report
Contractual date of delivery	:	2014-01-31
Actual date of delivery	:	2014-01-31
Lead beneficiary	:	LUH
Author(s)	:	Michelle Lindlar <michelle.lindlar@tib.uni-hannover.de> (LUH), Hedda Saemann <hedda.saemann@tib.uni-hannover.de> (LUH), Sebastian Ochmann <ochmann@cs.uni-bonn.de> (UBO), Ujwal Gadiraju <gadiraju@l3s.de> (LUH), Östen Jonsson <osten.jonsson@ldb-centrum.se> (LTU).

Responsible editor(s)	:	Michelle Lindlar <michelle.lindlar@tib.uni-hannover.de> (LUH).
Quality assessor(s)	:	Jakob Beetz <J.Beetz@tue.nl> (TUE), Martin Tamke <martin.tamke@kadk.dk> (CITA), Raoul Wessel <wesselr@cs.uni-bonn.de> (UBO).
Approval of this deliverable	:	Stefan Dietze <dietze@L3S.de> (LUH) – Project Coordinator
Distribution	:	Public
Keywords list	:	gap analysis, state of the art, digital preservation

Executive Summary

This deliverable identifies gaps in existing processes for the digital preservation of 3D objects. The gap analysis is approached through an in-depth analysis of two areas. One area is that of fundamental digital preservation tools and processes regardless of their content type. It describes processes and standards adapted by the global digital preservation community and implemented in archives of varying domains, e.g., archives dealing predominantly with e-publications as well as AV-archives. The second area is that of current existing processes for the digital preservation of 3D objects. It describes aspects and challenges which are uniquely tied to the long-term archiving process of this content-type and lists existing tools and standards. The gaps are identified through a comparison of the content type agnostic and the 3D-specific state of the art descriptions.

Table of Contents

1 Introduction

In dealing with digital preservation, the Reference Model for an Open Archival Information System (OAIS) is usually the first point of reference. It has given the preservation community a common vocabulary and foremost established a framework of concepts describing the processes needed to accept the responsibility of long-term stewardship for sustainability and accessibility of digital objects in the face of changing technology. However, as a reference model, the OAIS has its limitations as it can only deliver a high-level description of objects in the juxtaposition between producer, archive and consumer. Full lifecycle implications of the objects, as well as domain-specific needs, are out of scope for the OAIS.

Awareness of different risks associated with the long-term accessibility of digital information rose in particular in the mid- to late nineties. Reports such as that of the CPA/RLG (Commission on Preservation and Access / Research Library Group) “Task Force on the Archiving of Digital Information” demonstrated that the problem was now being addressed at the highest levels of the information services and cultural heritage domains. Around the same time the term “digital dark ages” [25] was coined and Jeff Rothenberg stated that “Digital objects last forever - or 5 years, whichever comes first” [38].

Thibodeau proposed in 2002 that a digital object consists of three layers: a physical, a logical and a conceptual layer. In digital preservation the properties of all three layers need to be considered and their relations to each other need to be understood [46]. In digital preservation discourse, the layers identified by Thibodeau have been addressed in “bit preservation”, “logical preservation” and “semantic preservation” [27].

In the process of maintaining the accessibility and understandability of an object over time, all three layers have to be taken into consideration. The lowest level - bit preservation - is largely content, domain and representation agnostic, meaning that no knowledge of the object’s format, information content or context in which the object was created in is required in order to address it. The usage scenario, however, may play a role in bit preservation as factors like consumer requirements may result in decisions regarding offline, nearline or online storage. The semantic preservation layer, on the other hand, focuses mainly on the long-term understandability of the content and captures information about the domain in which the object was created. The representation form of the object as well as the data stream underneath plays a minuscule role in semantic preservation.

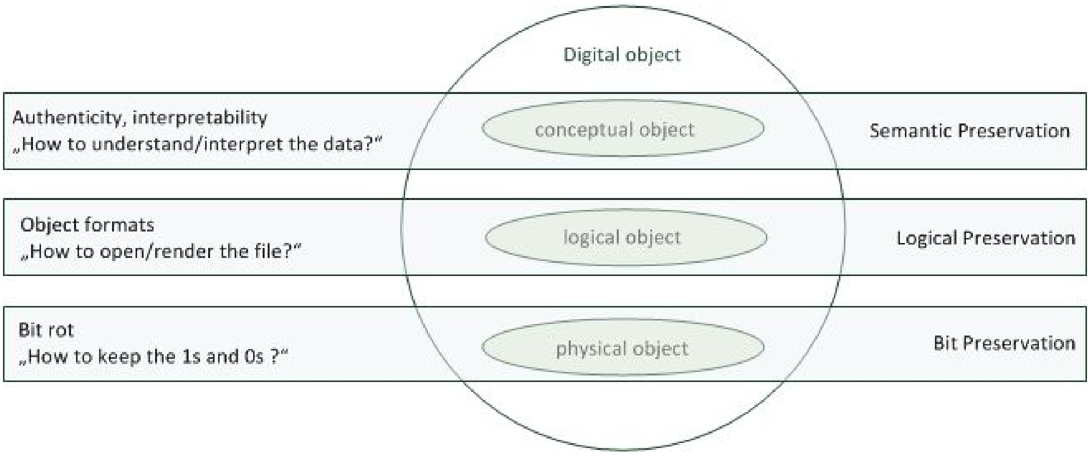


Figure 1: The three layers of a digital object

The focus of logical preservation, however, is clearly on the representation of the object, which needs to be in a form suitable for the content and accepted by the domain.

Based on the domain an object stems from or the usage scenario it is being archived for, material of the same content type may be treated differently on the layers of bit preservation, logical preservation and semantic preservation. 3D data, for example, is being used in various domains today, such as product development, archaeology, computer games or architecture. While the content type is the same for all domains, the objects vary in file format, accompanying metadata, environment they were created in and intended re-use. Every domain will have to address all three layers of the object, but chosen approaches will certainly differ in some of the processes.

Chapter 2 describes existing tools and standards in digital preservation on a content agnostic level. After a brief introduction of digital preservation processes in form of the Reference Model for an Open Archival Information System (OAIS), de-facto standards and existing best practises for bit preservation, logical preservation and semantic preservation are given. A chapter on metadata shows how the information gained in the different processes is captured and stored alongside the digital object to document provenance, authenticity, integrity and context. Domain and organization specific factors are an integral part of digital preservation processes. The section on organizational roles highlights the impact of the different stakeholders and gives insight into organizational processes relevant within the DURAARK scope.

While chapter 2 describes the state of the art of digital preservation at large, chapter 3 gives an insight into 3D content specific factors to be considered as well as into processes already in place. In a first step different projects and guidelines which are of relevance to the preservation process of architectural 3D data are explored. While section 3.2 briefly revisits bit preservation, section 3.3 is an in depth analysis of the two main file formats of the DURAARK project - IFC-SPF and E57 - in regards to logical preservation. The chapter covers the sustainability factors previously defined in 2.3.1 and tests existing digital preservation tools towards their support of the two file formats. Sections on semantic preservation, metadata and organizational preservation analyze domain specific standards and needs.

Chapter 4 describes the gaps identified by comparing the state of the art of digital preservation found in chapter 2 with the current state of 3D object preservation described in chapter 3. Each of the preservation processes previously described - i.e., bit preservation, logical preservation, semantic preservation, metadata and organizational roles - is analyzed in regards to implementation and knowledge gaps which are briefly listed.

2 Digital preservation - existing tools and standards

As described in the introduction, digital preservation research and practise first rose in the mid-1990ies. The OAIS reference model provided a basis for common understanding of concepts and vocabulary, which helped establish the research field of digital preservation further. European community funded digital preservation research activities started in the first years of the 21st century with projects like ERPANET (2001-2003), which established a network for digital preservation knowledge exchange and the DELOS (2004-2008) digital library reference model, which included preservation as a set function. As Strodl et al. [43] point out, early efforts in digital preservation were targeted towards simple textual documents and images.

The PLANETS project (2006-2010) developed first tools and frameworks supporting different preservation tasks, such as file format characterization, migration, emulation and preservation planning. The SHAMAN project (2007-2011) investigated preservation processes across distributed environments and was the first European project to include objects out of the engineering, more specifically the product-lifecycle-management domain. Strodl et al. further point out that main targets of current European research initiatives can be grouped into three areas: networking activities such as training, audit and certification (e.g., SHAMAN); applied research mainly dealing with scalable preservation as well as automation and decision support tools (e.g., SCAPE, ARCOMEM, ENSURE) and fundamental research dealing with interactive and embedded objects, ontologies, validation and preservation action quality assurance (e.g., LIWA, TIMBUS, SCAPE) [43].

The following sections describe the current state of preservation processes. The areas covered are in-line with a holistic preservation approach, covering all three layers of an object, as well as metadata and organizational roles in digital preservation.

2.1 The Reference Model for an Open Archival Information System (OAIS)

The “Reference Model for an Open Archival Information System (OAIS)” is a standard work describing components and services required within a long term archive. The archive itself is often referred to as an “OAIS” - an open archival information system. The reference model defines an OAIS as an **archive**, which has accepted the responsibility to **preserve** data for a **designated community**. Within this description, the following definition holds true for the key terminology [13]:

- an **archive** is not only software or hardware but a combination of an organization, people and systems
- to **preserve** means to store and to maintain the accessibility to information
- the **designated community** is a group of people identified by the archive as potential consumers.

The reference model was developed by the Consultative Committee for Space Data Systems (CCSDS). While the main stakeholders of the CCSDS are space agencies, the reference model was fast adopted by all domains dealing with long-term archival and has become a fundamental pillar of digital preservation research and practise. The first openly available version of the OAIS was the "Blue Book" (2001) which is identical with ISO 14721:2003. After an intermediate draft version in 2009 ("Pink Book") a revised version of the reference model was published by CCSDS as the "Magenta Book" in 2012. In the same year, the revision was also accepted as a new ISO standard revision (ISO 14721:2012).

The standard is to be understood as a framework, defining terminology and concepts for the description and comparison of preservation strategies. It does not include an implementation or design specification and explicitly states that implementations may choose to group the defined functionalities differently [13].

The reference model defines six functional entities within an OAIS:

1. *Ingest* provides services and functions connected with accepting the objects from an external or internal producer and preparing the information for archival storage and management, such as performing quality assurance or extracting descriptive information.
2. *Archival Storage* provides services and functions connected with storage, maintenance and retrieval of objects, such as managing storage hierarchy or refreshing storage media.
3. *Data Management* provides services and functions connected with populating, main-

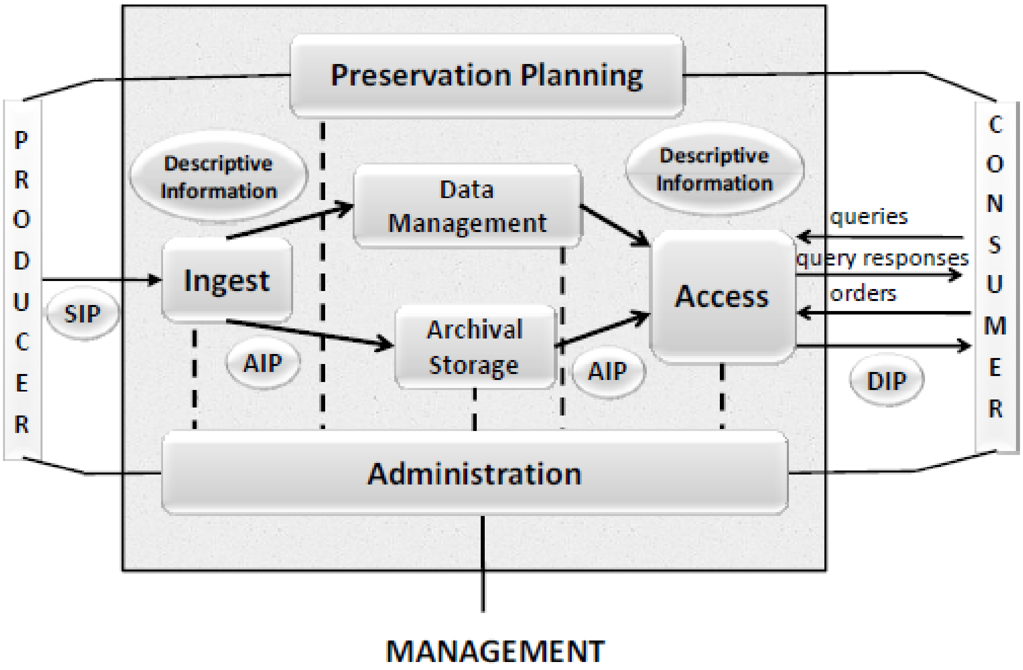


Figure 2: Functional entities of the OAIS [13]

taining and accessing descriptive information on the objects, such as ensuring that new information is loaded into the database.

4. *Administration* provides services and functions needed for the overall operation of the system, such as maintaining configuration of hardware and software or monitoring activities.

5. *Preservation Planning* provides services and functions needed for awareness of changing technology and community requirements, such as monitoring, evaluation and policy development.

6. *Access* provides services and functions needed in enabling the consumer to locate and receive the archived objects, such as coordinating delivery and enforcing access limitations.

In addition to the functional entities, figure 2 shows the flow of information packages within the system. The model has defined three different information packages for three different stages of the information flow: the submission information package (SIP), the archival information package (AIP) and the dissemination information package (DIP).

The information packages vary in required content. While a producer may choose to distribute minimal and detailed description information about an object in separate SIPs, an OAIS may decide that full preservation information including fixity information may not be of relevance to the designated community and does therefore not need to be part of the DIP. In differentiation to the SIP and the DIP the Archival Information Package AIP needs to function as “a container that contains all the needed information to allow Long Term Preservation and access to Archive holdings” [13]. The structure of an AIP, which is therefore the most exhaustive one, is shown in figure 3.

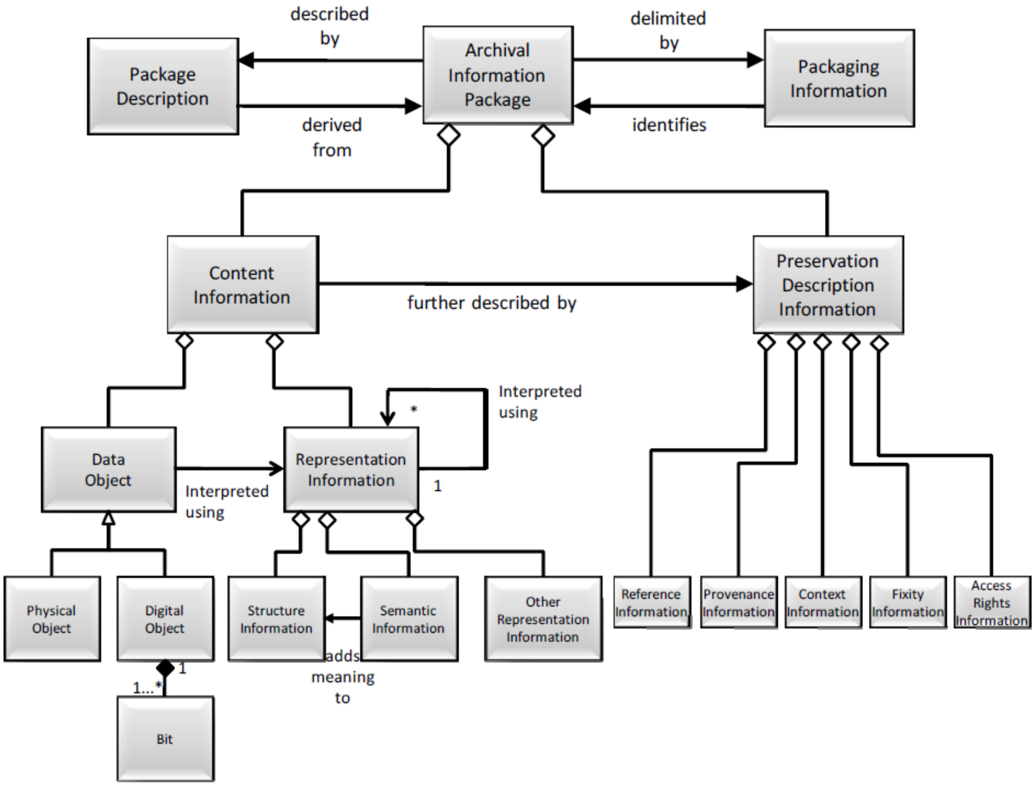


Figure 3: Detailed View of an Archival Information Package [13]

The archival object itself is the “digital object”. To ensure the understandability of the object over the course of time, it needs to be accompanied by “representation information”. This information shall ensure the understandability on two levels: on a semantic level, ensuring that the content and context can be understood and on a structural level, ensuring that the object can still be rendered/represented in the intended way. The archival object as well as the representation information are defined as “content information” within the OAIS reference model.

Next to the content information, an AIP shall contain a second information type - the “Preservation Description Information” (PDI). PDI is additional information needed for preservation purposes, such as provenance, access rights or reference via external identifiers. Furthermore, as maintaining the object over time inevitably means changing the object, e.g., in form of migrating the object at a structural or semantic level, the authenticity of the object can only be maintained if a documentation of these changes is stored alongside the object in the PDI.

The likelihood of change as part of the preservation process brings another requirement: the definition of those characteristics of an object which need to be preserved over time. The reference model lists these characteristics as “Transformational Information Properties” which are “[...] regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately preserved information content” [13]. Examples for such properties are factors describing appearance or behavior. When transforming an object containing the periodic table, for example, the block layout is a significant characteristic which needs to be kept, whereas the font-size may be irrelevant. Transformational Information Properties are also, as the standard points out, known as significant properties (see chapter 2.6.2).

The term “OAIS compliance” is frequently used in the description of implemented archives. But what exactly is OAIS compliance? The reference model itself lists the following criteria for conformity [13]:

- the basic information model describing the concept of the information packages (including content information and PDI) as well as the producer and consumer interaction should be supported
- the OAIS shall fulfil the following responsibilities
  - negotiate for and accept appropriate information from producers
  - obtain a sufficient level of control over the information to ensure preservation
  - determine designated community and define knowledge that can be assumed for the community

- policies and procedures need to be defined, documented and followed; these should cover procedures for the case of the demise of the archive
- the preserved information needs to be made available to the designated community either as copies of the original or as new representations with traceable changes to the original

The reference model neither requires nor defines how those methods are to be implemented on a technological or organizational level.

2.2 Bit Preservation

Bit preservation is the basic layer for digital preservation as shown in figure 1 in the introduction section. Preservation activities at this level shall ensure the integrity of the sequence of the code (the “1s and 0s”) over time and are therefore the prerequisite for any following preservation activities.

The bit level of a digital object is put at risk by technologically intrinsic risks - e.g., partial or complete media failure, extrinsic risks - e.g., operator error or data abuse, as well as by risks derived from technological progress - e.g., data carrier obsolescence as in the case of the 8-inch floppy disk. In preservation practise, these risks are met through different actions including object replication across multiple storage systems, ideally of different type, through regular replacing or refreshing of the storage systems and through regular auditing of the object copies to detect damages and trigger repair. The auditing of objects across several storage systems is typically conducted through the generation, re-generation and comparison of checksums.

Factors such as disaster recovery shall not only cover the case of technological failure, but also the case of natural disasters or loss through, e.g., fire. A geographical spread of storage systems should be considered and transparent policies should be in place on the organizational level.

2.3 Logical Preservation

While bit preservation addresses the object at the very basic layer, logical preservation addresses the file format encoding of the object. Preservation activities at this layer shall ensure the ability to render the object and maintain accessibility over time.

As in the case of bit preservation, different risks exist on the object's logical layer. The possible chain of dependencies connected to the rendering process poses a major threat to logical preservation: an object of a certain format depends on rendering software, which may in return depend on an operating system or certain configurations or packages, which may depend on hardware. Obsolescence of any software or hardware which the rendering process depends on, poses a threat to the entire rendering process. Furthermore, malformed objects which do not comply completely to the file formats standard may not be renderable with every software supporting the file format. This will especially become problematic in a future scenario, where rendering software may have to be reengineered based on the file format's specification. In order to evaluate adherence to file format specification, the specification needs to be available and the format needs to be open to tool inspection - for proprietary file formats those two factors are unfortunately often not the case.

In developing digital preservation strategies to meet the constant change imposed on digital objects and their environments, file formats chosen for archival purposes need to be carefully evaluated. The following chapter will define sustainability factors for file formats, while chapters 2.3.2, 2.3.3 and 2.3.4 will look at best practise digital preservation processes for tool based inspection and analysis of a digital object at the logical preservation layer.

2.3.1 File Format Sustainability

When choosing a format for long-term digital preservation, a number of factors must be considered to ensure that the format is as long-lived - as sustainable - as possible. A number of long-term data stewards will only accept file formats into their archive which they deem sustainable by today's knowledge and normalize file formats not suited for long-term archiving to sustainable target formats suited for the respective content type.[1] Other institutions will include any file formats in their digital preservation system but only guarantee full preservation activities, including logical preservation, to file formats deemed sustainable.[2]

But what makes a format sustainable? This chapter puts forward requirements for file format sustainability, based on recommendations defined as part of the InterPARES project [34], by The National Archives (TNA) UK [11] and the Royal National Library of the Netherlands [37]. For each of the six main categories identified as sustainability characteristics, concrete factors are defined which file formats can be easily checked against.

1. **Disclosure**

   To really understand and interpret a logical format it is necessary to have an understanding of its design and structure, how the format stores the bit-stream. Knowledge of the file formats inner structure and syntax is necessary for a number of preservation activities, such as tool development, e.g., for technical metadata extraction or migration, for error checking and for the reconstruction of rendering software if the original software is for instance no longer available. Without this knowledge, the file is just a combination of ones and zeros, lacking logical meaning and preservation activities beyond bit preservation are almost impossible to achieve.

   Sustainability factors:

   - well documented and complete specifications
   - public (open) specifications
   - format specifications should be stable and - if changes occur - backward compatible

2. **Internal technical characteristics**

   This category looks at the technical mechanisms that affect the format's internal structure, such as encryption. Digital Rights Management (DRM) copy protection

[1] See for example the preservation file format table of the National Archives of Australia: `http://www.naa.gov.au/Images/Preservation-File-Formats_tcm16-79398.pdf`

[2] See for example file format recommendations of Purdue University: `https://purr.purdue.edu/legal/file-format-recommendations`

can significantly hinder preservation processes on all levels - from plain bit preservation backup practises to file format validation methods. To simplify maintenance, a format's complexity should meet the intended functionality and ideally internally support preservation processes through error-detection.

Sustainability factors:

- free from encryption
- free from Digital Rights Management (DRM) copy protection
- complexity should meet the intended functionality and not be over-specified
- error-detection included in format

3. **External technical characteristics**

   As mentioned in 2.3 dependency chains of software or hardware combinations pose a high threat to the preservation process as the availability and sustainability of every part of the dependency chain has to be ensured for successful rendering. Fewer dependencies on specific hardware or software therefore mean higher sustainability of the file format.

   Sustainability factors:

   - independent of hardware
   - independent of physical medium
   - independent of specific software or operating system
   - independent of external information

4. **Format Acceptance**

   A wide spread acceptance of a file format usually goes hand in hand with extended tool support. A typical example for this is the PDF file format which has been widely embraced as an access, but also as a preservation format for textual materials.[3] For some content types acceptance on a global level may not be possible, as the content representation form is highly specific to a certain domain. Acceptance

[3] For archival purposes PDF/A family formats are preferred. See for example the "local use" description for PDF at the Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml

should therefore be checked on several levels - globally, within one domain, within several domains. In this context standardization can be seen as a solid indicator for file format acceptance.

Sustaninability factors:

- support through several software manufacturers
- embraced / popular with industry
- used by several domains
- standardised (ISO, SIS, etc.)

5. **Patent**

   Patents can affect the usage as well as the maintenance of the digital format in the archive. The existence of a patent may prevent the future creation of "open source" software but also the usage and acceptance of the file format. It is important to note that patents may pertain to the entire format, but also only to algorithms used within the format - both can lead to problems in the preservation process. A well-known example for such a case is the gif file format, whose usage faced problems when the UniSys cooperation started to charge fees for the LZW compression algorithm between 1994 and 2003 [26].

   Sustainability factor:

   - free from patent / licensing costs

6. **Logical Structure and Transparency**

   A logical and transparent structure of a file format includes a clear differentiation between a header, which typically includes some information about the data stream, and the data or "payload" sector, which includes the actual data stream. An example for such a clear structured format is the RIFF (Resource Interchange File Format) based WAVE audio format or the image format TIFF (Tagged Image File Format), where the header may include information such as the payloads encoding (wave) [29]. Uncompressed payloads allow for a direct analysis of the data stream with simple external tools - such as a TIFF analysis with a hexeditor. For other formats analysis tools may already be available in the user or the digital preservation community. An example for such a tool availability is the ExifTool which can extract and manipulate metadata included in a variety of still image formats.[4].

Sustainability factors:

- existing methods for validation of file structure
- self-documented format, containing i.e., metadata such as information about the producing application
- the file's content is transparent for "simple" tools
- standard or simple representation of the data in the file (e.g., human readability)

2.3.2 Identification

In order to address risks at the logical preservation layer, an exact identification of the object's file format is a necessary first step. Operating systems usually rely on file extension or mime type for file format identification. As this is information which can easily be manipulated, digital preservation tools usually take a more forensic approach to file format identification. A common approach is comparing parts of the digital objects to file format patterns stored in databases [1].

Widely used file format identification tools in preservation practise are DROID[5], the UNIX file utility / the libmagic library[6], FIDO[7] or the closed source TrID file identifier[8]. The tools differ in their methods, the number of formats they support as well as in their correct identification of certain formats. Depending on the intended usage one tool might fair better than another, for example if processing time is a relevant criteria. Furthermore, a number of file formats are not supported by any of the available tools [48].

PRONOM[9] and the UDFR (Unified Digital Format Registry)[10] are two registries which

[4] A list of the metadata formats currently included is available on the tool's website: http://www.sno.phy.queensu.ca/~phil/exiftool/

[5] http://droid.sourceforge.net/

[6] http://sourceforge.net/projects/libmagic

[7] http://www.openplanetsfoundation.org/software/fido

[8] http://mark0.net/soft-trid-e.html

[9] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[10] http://www.udfr.org/

aid the identification process by maintaining information about a wide variety of file formats such as information about the vendor and patents, links to documentation and related file formats. PRONOM further assigns an unique identifier to each file format - the PUID (PRONOM Unique IDentifier).

2.3.3 Technical Metadata Extraction

While file format identification is the first necessary step in preservation processes on a logical level, further information about the file format needs to be gathered in a second step. The encoding used in a format, e.g., PCM (pulse code modulation) or mp3 in a WAVE container, the quality level of the information, e.g., a high resolution in a TIFF file or whether fonts are embedded or linked to a PDF file have an impact on preservation decisions. Furthermore the object may contain metadata documenting the provenance or creation process, e.g., in the form of an embedded author tag, a time stamp or information regarding the creating application. A number of tools exist for the process of technical metadata extraction. While few tools - namely jhove[11], jhove2[12], Apache Tika[13] and the Metadata Extraction Tool[14] - support tools of varying content type, the majority of tools either support one file format family (e.g., the pdftk toolkit for PDF formats[15]) or several formats of the same content type (e.g., Mediainfo for audio and audiovisual materials[16].

2.3.4 Validation

Like file format identification, file format validation is a central aspect of the preservation of objects at a logical level. File format validation checks standard and schema conformity of objects. The output of validation components may be broken down into statements whether an object is “well-formed” and whether an object is “valid”. Well-formedness refers to the low level syntax of an object. For example, XML files are considered well-formed when the object adheres to the syntax rules specified in the XML specification. These syntax rules define, e.g., that a document has a single root element, that each

[11] `https://sourceforge.net/projects/jhove/`

[12] `http://jhove2.org/`

[13] `http://tika.apache.org/`

[14] `http://meta-extractor.sourceforge.net/`

[15] `http://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/`

[16] `http://mediainfo.sourceforge.net/en`

element must have a closing tag and that elements are properly nested. The XML specification does not, however, pre-define a set of tags or attributes. In the case of XML this may be done via a schema. The conformity check against a schema determines whether an XML is “valid”.

Based on this, an object can only be “valid” when it is “well-formed”. Lack of well-formedness and/or validity have an impact on preservation capabilities. Examples for file format validation tools include the W3C Markup Validation Service for DTD-based formats like HTML and XHTML[17], jpylyzer for the JPEG2000 file format[18] or jhove for the format families PDF, AIFF, GIF, JPEG, JPEG2000, TIFF and Wave[19].

2.4 Semantic Preservation

The OAIS describes semantic information as “the representation information that further describes the meaning beyond that provided by the structure information”’ giving examples such as the language in which a text is written [13]. While the language a text is written in does of course not change over time, the knowledge of a language might change over time. An example for a concept which changes faster is that of a price list in a document, where the value of a price will change with inflation or a currency will change such as in the case of the introduction of the Euro. Without capturing the necessary knowledge to interpret the information on an intellectual level, the original meaning will be lost over time. Schlieder describes this risk as ‘cultural ageing’, stating that: “The corresponding documents are no longer retrieved, the data is no longer used in inferences. Knowledge about the semantics of digital records may persist for a while after the community loses interest in their content. However, as the semantic knowledge is not maintained and transmitted any more, its loss is almost unavoidable” [40].

Semantic preservation is so far the least addressed of the digital object layers shown in figure 1. Strodl et al. identify semantic preservation as a main future research area of digital preservation, advising a close cooperation between semantic web and digital preservation experts [43].

Up to now semantic technologies in digital preservation have mainly been applied to

[17] http://validator.w3.org/

[18] http://www.openplanetsfoundation.org/software/jpylyzer

[19] http://sourceforge.net/projects/jhove/

further logical preservation efforts, such as in the P2 Registry developed by the University of Southampton [45]. First efforts in semantically enriching archives with social web data and thus moving towards a formation of semantic categories in a preservation approach are currently being undertaken in the EU FP7 project ARCOMEM[20].

While ARCOMEM is specifically looking at data from the social web, an abundance of data on the web exists which can be used for semantic enrichment. The Linked Open Data (LOD) cloud offers a vast amount of data of both domain-specific and domain-independent nature. Linked data describes a method for publishing structured data so that it can be interlinked and become more useful. Such interlinking enables data from a wide variety of sources (e.g., geonames.org, DBpedia) to be connected and queried. While the enrichment with this data is a first step towards semantic preservation, the nature of the distributed, inter-linked sources of data also introduce a new set of significant challenges from a preservation point of view.

A first challenge is naturally the question of identifying datasets suited for enrichment. This encompasses the discovery and analysis of relevant datasets and (or) endpoints, apart from dataset profiling and description. Dataset discovery can be riddled with obstacles like ill-described datasets, with little or no structured information about the quality and coverage of the dataset as well as a lacking understanding of how persistent the dataset is. This calls for methods for data curation and dataset profiling. There has been a fair amount of research in this realm. Rula et al. investigate the characterization and availability of temporal information in linked data at a large scale [39]. The authors of [19] introduce a processing pipeline to automatically assess, annotate and index available linked datasets. The generated profiles embed datasets into an interlinked data-graph of datasets based on shared topics and vocabularies. Some earlier works address related issues [15][44], such as schema alignment and extraction of shared resource annotations across datasets.

The distributed nature of LOD as a potential semantic enrichment candidate requires a number of preservation activities addressing the stability of the dataset chosen. Due to the inherent nature of linkage in the LOD cloud, changes with respect to one part of the LOD graph are propagated throughout the graph. Hence, measuring the impact of a change in one dataset (entity) on other datasets (entities) within the LOD graph is crucial. Tracking evolutionary changes in linked datasets is a relatively new realm of research.

[20] `http:\\www.arcomem.eu`

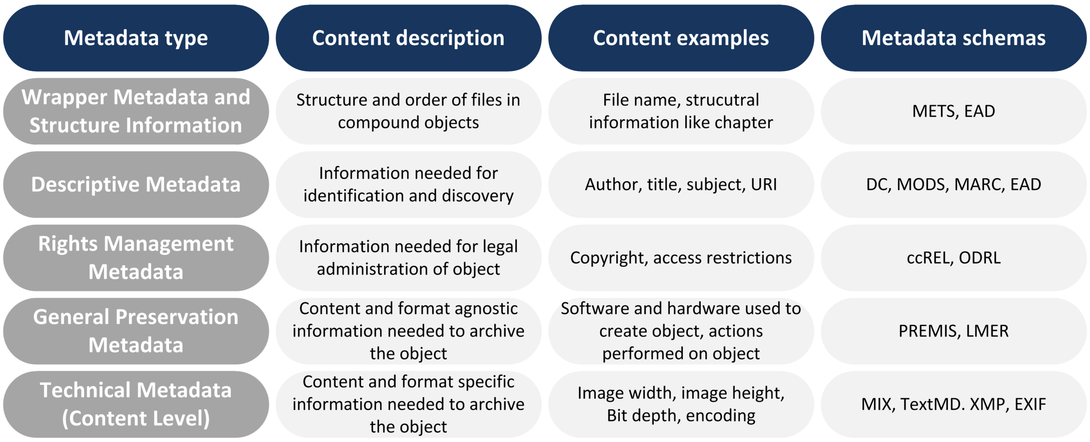
Käfer et al. present initial results from the Dynamic Linked Data Observatory: a long-term experiment to monitor a two-hop neighborhood of a core set of diverse linked data documents [24]. The authors investigate the lifespan of the core set of documents, how often they stay on-line or go off-line and how often they change. Furthermore, they delve into how links between dereference-able documents evolve over time. An understanding of how links evolve over time is essential for traversing linked data documents, in terms of reachability and discover-ability. Ntoulas et al. [31] discovered that hyperlinks in HTML documents tend to be more dynamic than other forms of content.

2.5 Metadata

As described in chapter 2.1, the information model of an OAIS shall include certain information about the archival object within its information packages. In order to fulfil the task of preservation, the archive needs to have a full understanding of the object it wants to preserve and the designated community it wants to preserve the object for. This includes knowledge about technical and contextual criteria - only if we understand how the object can be rendered technically and interpreted semantically can we guarantee accessibility and understandability over time. This information is captured in metadata which is, as the National Information Standards Organisation points out the “key to ensuring that resources will survive and continue to be accessible into the future” [30].

As the OAIS reference model does not define how the information is captured, several implementation approaches and standards have been established. In general, metadata can be captured embedded within an object as well as in external files storing the information. In digital preservation practise, common practise is to extract metadata where possible and to store it in separate files and/or databases to ease search and retrieval inline with the OAIS data management entity [30]. Metadata in archival systems can be divided into three functional categories: “descriptive metadata” “structural metadata” and “administrative metadata” [30]. While “descriptive metadata” contains information needed for identification and discovery and “structural metadata” deals with the organization of multiple files into a meaningful object, “administrative metadata” sums up a number of functions, such as rights management, provenance and technical metadata describing, e.g., the quality of the object.

Often, metadata standards cover different areas of categories. This is frequently the case



Metadata type	Content description	Content examples	Metadata schemas
Wrapper Metadata and Structure Information	Structure and order of files in compound objects	File name, strucutral information like chapter	METS, EAD
Descriptive Metadata	Information needed for identification and discovery	Author, title, subject, URI	DC, MODS, MARC, EAD
Rights Management Metadata	Information needed for legal administration of object	Copyright, access restrictions	ccREL, ODRL
General Preservation Metadata	Content and format agnostic information needed to archive the object	Software and hardware used to create object, actions performed on object	PREMIS, LMER
Technical Metadata (Content Level)	Content and format specific information needed to archive the object	Image width, image height, Bit depth, encoding	MIX, TextMD. XMP, EXIF

Figure 4: Metadata types, content and standards in archival practise

for metadata standards developed within a domain, where a standard was developed to describe a specific intellectual content type or logical content group in an all-encompassing way. Examples for this are the DDI[21] (Data Document Initiative) standard for social science data, TEI[22] (Text Encoding Initiative) for linguistics data or MPEG-7[23] for AV material.

As digital preservation activities in the past were largely driven by large cultural heritage institutions [43] who hold archival responsibility for material of varying content type and domain origin, several domain independent de-facto standards have been established. Figure 4 shows the main metadata types addressed in digital preservation practise. The de-facto standards listed are domain and organization agnostic. The only content dependant standard is technical metadata, which describes content and format specific information needed for long-term preservation.

2.5.1 Preservation Metadata

PREMIS[24] (PREservation Metadata: Implementation Strategies) is recommended by the OAIS as a standard for the submission of digital metadata about the object to an archive [13]. PREMIS understands preservation metadata as data drawn from different information sources (see figure 5).

[21] http://www.ddialliance.org/

[22] www.tei-c.org/

[23] http://mpeg.chiariglione.org/standards/mpeg-7

[24] http://www.loc.gov/standards/premis/

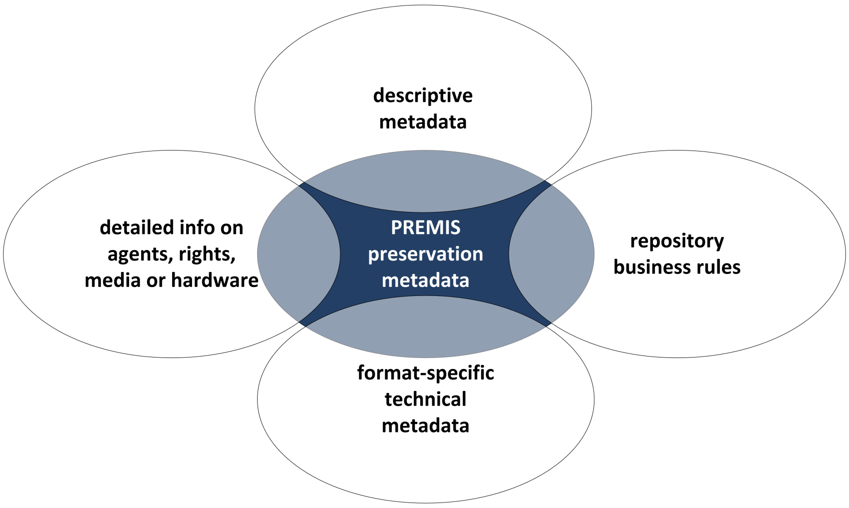

descriptive metadata
detailed info on agents, rights, media or hardware
PREMIS preservation metadata
repository business rules
format-specific technical metadata

Figure 5: PREMIS information sources (based on Caplan [12])

The standard consists of a data dictionary, which is regularly being revised based on community input and is currently available in version 2.2, as well as of an XML schema. A draft OWL ontology of the data dictionary version 2.2 is also available via the PREMIS website. Within PREMIS information such as an object's fixity information, significant properties, extracted technical metadata and file format information is captured. To support further granularity, PREMIS allows extensibility of several semantic units, one being the objectCharacteristics unit [35]. ObjectCharacteristics contains the aforementioned content specific technical metadata about a file. The PREMIS data dictionary describes technical metadata as information which "describes the physical rather than intellectual characteristics of digital objects" [35]. The standard further states that as technical metadata is highly dependant on the nature of the content and on the capabilities of file formats, the development of corresponding objectCharacteristic parameters should "be left to format experts" [35] who may use external technical metadata schemas in the extendable objectCharacteristics semantic unit.

An example for such a content specific standard for technical metadata is MIX - the NISO metadata standard for still images [25]. Any still image object may be described using MIX metadata. The schema captures information in 5 sections [2]:

- **Basic digital object information** such as object identifier, file size, byte order and compression information (schema, ratio).

[25] http://www.loc.gov/standards/mix/

- **Basic image information** which consists of basic image characteristics such as image width, height, color space and color profile - as well as a few special format characteristics defined for the jpeg2000, MrSID and Dejavu formats.
- **Image capture metadata** consisting of source information and general capture information as well as specific capture information for the capture sources scanner and digital camera.
- **Image assessment metadata** such as spatial metrics, detailed information about the color encoding like primary chromacities and target information.
- **Change history** such as processing software and processing rationale.

Technical metadata is often used synonymously with significant properties. As mentioned in 2.1 significant properties define criteria which should be preserved across successive cycles of preservation processes. Significant properties often contain technical metadata which describes the quality, structure or behavior of an object. However, they are not exclusively derived from technical metadata, as described in 2.6.2.

2.6 Organizational Roles in Digital Preservation

The OAIS reference model includes the organizational roles of preservation from the get-go, describing an OAIS as “an archive, consisting of an organization, [...] of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community” [13]. Organizational and technological processes must therefore go hand in hand when establishing digital preservation processes.

The OAIS functional entities described in chapter 2.1 all contain technological and organizational processes. The following section will give a brief insight into organizational factors which will play a role within the DURAARK project: lifecycle models, preservation planning and significant properties. While many more organizational aspects exist - such as trustworthiness, sustainability of the archive itself, certification processes and policies, to just name a few - those are not within the scope of the DURAARK project.

2.6.1 Lifecycle models

In order to preserve objects for a designated community, knowledge about the context in which the object was created in, as well as knowledge about the intended use and re-use is necessary. The impact of actions on an object's curation and preservation process is best understood when seen in the full context of a lifecycle view.

Various domain and content-specific data lifecycle models exist, such as the DDI Combined Life Cycle Model for social, behavioral and economic sciences or the I2S2[26] (Infrastructure for Integration in structural Sciences) Idealized Scientific Research Activity Lifecycle Model, which is tailored towards the needs of data from structural sciences such as chemistry.

A current study of lifecycle models conducted by Ball [6] shows a high number of lifecycle models in connection with research data management. Ball points out that curation and preservation actions can be made easier when planned and prepared for in advance - a process in which lifecycle models are a helpful communication and planning tool.

The DCC (Digital Curation Centre) Curational Lifecycle Model is a domain agnostic description of the lifecycle of an object from its conceptualization to its continuous use, disposal or re-use and transformation which leads to a new object. The model is a generic and high-level one. The authors of the model point out that it may be used in conjunction with further reference models, frameworks or domain-specific tools and standards to take more granular approaches [22].

The DCC model is a planning tool for producers, users and data custodians. It consists of a number of sequential actions which describe the full lifecycle and can either fall in the category of curational actions or preservation actions. In addition, the model defines three actions which shall accompany an object throughout the entire lifecycle:

1. **Management of description and representation information:** The DCC model defines "description information" as administrative, descriptive, technical, structural and preservation metadata, which shall be assigned and managed using appropriate standards (see also 2.5)."Representation information" is information needed to understand and render the objects and the metadata over time - this

[26] http://www.ukoln.ac.uk/projects/I2S2/

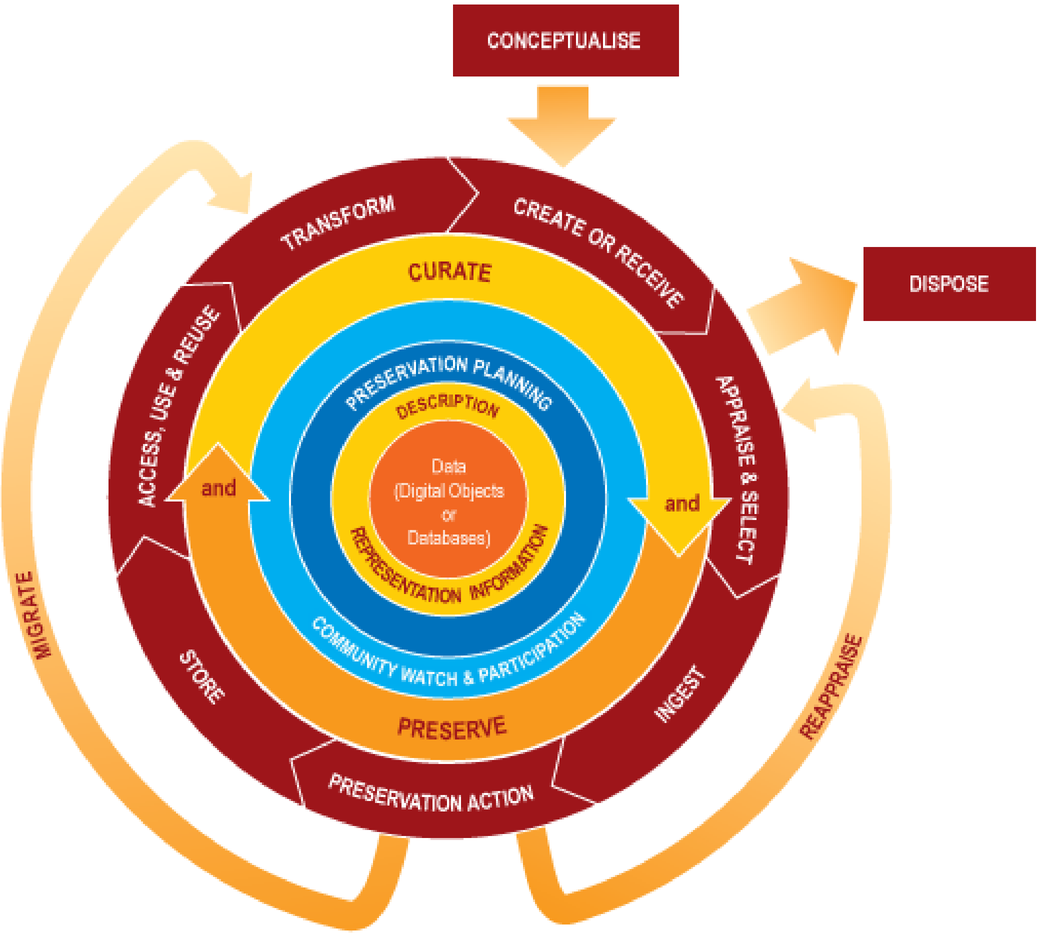


Figure 6: Preservation - Curation Lifecycle Model [22]

can, for example, be descriptions of the file format or metadata schema which is stored in the archive's knowledge base [22].

2. **Community Watch and Participation:** As a full lifecycle action, community watch includes all stakeholders involved: producer, consumer and custodian. The respective community should be monitored for changing expectations, emerging standards and best-practises. Participation describes the active involvement in furthering and development of standards, tools and suitable software.

3. **Preservation Planning:** The process of continuous re-evaluation of preservation measures throughout the entire digital object's lifecycle. Preservation Planning is described further below.

2.6.2 Significant Properties

The DCC curational lifecycle model shows that there are three stakeholders in the preservation process: the producer, the custodian and the consumer. A key task of digital preservation is defining the requirements which the stakeholders have in the preservation of an object. In the face of rapidly changing technology the preservation of objects is inevitably connected with having to change the object itself -as in the case of migration- or the environment -as in the case of emulation.[27] Maintaining every aspect of an object over the course of these changes is a costly, infeasable and also sometimes unwanted process, as new technology enables new usage scenarios which the object in its original preservation form may not be suited for. It therefore becomes essential to define those characteristics, which are essential for the continuous process of guaranteeing the object's accessibility, usability and meaning. This process serves two purposes:

1. **A common understanding of what is important is reached, considering the requirements of producers, custodians and consumers.** As the characteristics are based on the requirements of the stakeholders they are subjective. It is furthermore understood that they may change over time and should therefore be re-evaluated regularly.
2. **The defined characteristics shall serve as verification measures to check whether the requirements have been kept across preservation action.** This is supported through a formalized approach of capturing the requirement: characteristics consist of a property or facet with a respective value.

In digital preservation discourse different terminology has been used for this concept, such as "significant characteristics", "significant properties", "essence", "aspects" or "transformational information properties". Research work on the concept was conducted as part of projects like Cedars[28], CAMiLEON[29], InSPECT[30] and PLANETS[31] with further

[27] While emulation is based on the imitation of original environments which may include any combination of rendering software, operational system and hardware, e.g., the I/O devices used with the emulated environments are usually those of the present. This may significantly change the perception of the emulated object. An example for this is the rendering of digital art out of the CRT-era on present day LCD screens.

[28] http://www.ukoln.ac.uk/metadata/cedars/

[29] http://www2.si.umich.edu/CAMILEON/about/aboutcam.html

[30] http://www.significantproperties.org.uk/

[31] http://www.planets-project.eu/

work done at the National Archives of Australia[32]. The SCAPE project[33] is currently working on developing methods to include these requirements in a human and machine readable control language which can be passed from policies to the preservation planning process [41].

Dappert and Farqhuar developed a concrete definition of the concept within the PLANETS project, describing significant characteristics in their role as requirements:

"Requirements in a specific context, represented as constraints, expressing a combination of characteristics of preservation objects or environments that must be preserved or attained in order to ensure the continued accessibility, usability and meaning of preservation objects, and their capacity to be accepted as evidence of what they purport to record" [14].

Characteristics can therefore stem from three classes: the preservation object, the environment and the preservation action. Properties shall be defined for each of those classes [14]. A list of possible properties is currently being developed as part of the SCAPE control language [41]. Object based characteristics are typically based on technical metadata standards for the respective content type, such as AES metadata for audio[34] or MIX metadata for still images[35], or on metadata available to singular formats or groups of formats, such as EXIF data for JPEG, TIFF and RIFF WAV[36] or the bext chunk for Broadcast Wave (BWF) files[37].

2.6.3 Preservation Planning

While the DCC curation lifecycle model sees "preservation planning", "community watch and participation" and "manage descriptive and representation information" as separate activities, the OAIS reference model has a broader definition of the functional entity preservation planning:

"*The OAIS functional entity which provides the services and functions for monitoring the environment of the OAIS and which provides recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable*

[32] http://www.naa.gov.au/

[33] http://www.scape-project.eu/

[34] http://www.aes.org/publications/standards/search.cfm?docID=84

[35] http://www.loc.gov/standards/mix/

[36] www.cipa.jp/std/documents/e/DC-008-2012_E.pdf

[37] http://tech.ebu.ch/docs/tech/tech3285.pdf

by, and sufficiently usable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete" [13].

First work in preservation planning methodology was conducted as part of the DELOS project. The PLANETS project[38] further refined this methodology and developed the preservation planning tool "Plato" as a key outcome of the project. Plato functions as a decision support tool and allows the formulation, testing and evaluation of a preservation plan. A basic overview of the Plato workflow can be seen in figure 7.

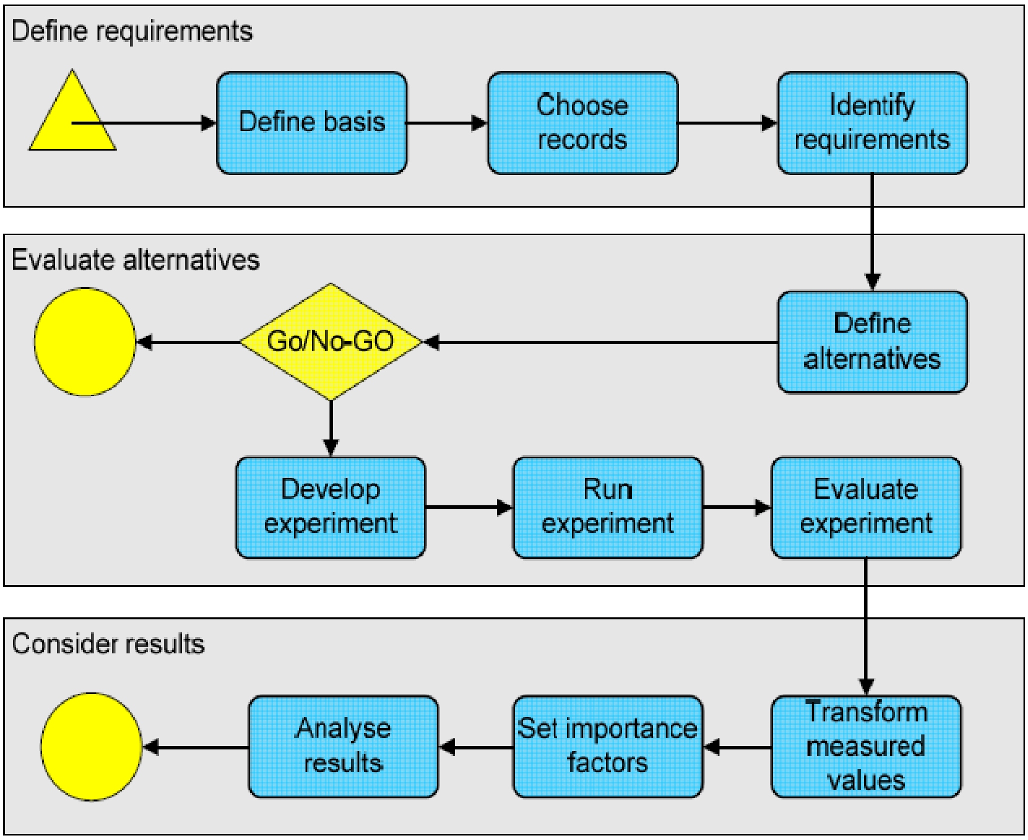


Figure 7: Plato preservation planning workflow [8]

Within Plato a collection is described through requirements and constraints, the aforementioned significant properties. As mentioned above, input factors for the requirements can be various sources, such as policies, legal constraints, organisational requirements, user requirements or characteristics of the digital object. A representative sample for the

[38] http://www.planets-project.eu/

collection is formed, a planned action chosen and run against the test set in an experiment. The outcome of the experiment is evaluated against the formulated requirements for the collection. The evaluation forms the basis on which an institution can make the decision on whether a preservation action should be taken on the collection or not.

3 3D preservation - existing tools and standards

The digital preservation approaches described in the previous chapter are widely accepted as good practise and may in theory be implemented for every type of content. However, tools and processes need to be in place for the content type which the processes are supposed to be leveraged on. As such, identification tools need to support file formats and technical metadata extractors content types; metadata standards must allow for the capturing of domain specific context information; preservation planning processes must have identified community sources relevant for the information to be archived. Furthermore, new content types may introduce new questions to the preservation process. An example for this might be a new risk which is innate to an object's feature. This chapter will analyze existing processes as well as special requirements for 3D architectural data. It will start out with an analysis of related projects and guidelines to have a reference basis of state of the art projects in 3D preservation. The following subchapters will follow the same outline as the one used for chapter 2.

3.1 Projects and Guidelines

Only two projects could be identified which specifically targeted 3D architectural data: MIT FACADE and DEDICATE. A third project of high relevance is the 3DCOFORM project - while the preservation of 3D architectural data was not the main focus of the project, it was covered within the project. Another project, the ongoing LOTAR project, addresses the long term archiving of 3D and product data management (PDM) data from the aerospace and defence industry, building on STEP application profiles.

Only one guideline pertaining to three-dimensional data could be identified: the London Charter. It identifies principles underlying the employing of three-dimensional visualisation technologies in heritage research and distribution and is described in a separate subchapter.

Other research of relevance to the DURAARK project objectives can be grouped in 4 categories: related virtual heritage projects, related preservation process projects, related linked data projects and related guidelines and strategies which do not stem from research projects but from research conducted within an institution. A respective subchapter for each category summarizes relevant projects pertaining to the category.

3.1.1 MIT FACADE

One of the main recent projects in 3D long term preservation was MIT's FACADE (Future-proofing Architectural Computer-Aided Design).[39] Running from 2006 to 2009 FACADE aimed at developing methods and best practices for the capture, description, management, preservation and availability of architectural CAD models. The project objectives were:

- to analyse the proprietary and short-lived CAD formats
- to supply the format identification to the digital format registry PRONOM[40]
- to develop guidelines for process documentation and annotation of CAD files
- to operate their ingestion, management, preservation and dissemination within a digital archive system.

Other digital material (e.g., images, specifications) accumulated during the building process was also considered. The project worked with limited test data collections (from merely 20 000 up to 100 000 files per collection) of four major projects covering different CAD modelling tools and file systems. Tools developed within FACADE[41] include modules for MIT's DSpace based digital archive system, which was enhanced for 3D data. In order to describe the relations of the annotated models to other building data, the Project Information Model (PIM) was created. This information ontology provides each file with a contextual structure based on properties which complied with metadata standards from the art and architecture library community (e.g., schemas: Cataloguing Cultural Objects (CCO) and Categories for the Description of Works of Art (CDWA)).

A "project" entity was established for placing the cataloguing metadata. The archiving workflow set up by FACADE covers the process from the receipt of a file to its release via the end user interface. Other workflows – preview workflow, post-publish workflow and license workflow – remained rudimentary. The public user interface as a platform

[39] http://facade.mit.edu/. FACADE was funded by the U.S. Institute of Museum and Library Services (IMLS). Partners involved were the MIT Libraries and MIT School of Architecture and Planning.

[40] PRONOM was developed by the Digital Preservation Department of the UK National Archives. See www.nationalarchives.gov.uk/PRONOM/

[41] Software developed within the FACADE project has been partly included in the subsequent DSpace version and partly archived in the MIT Libraries' software repository available on request under an open source software license.

for the end users was based on the technology for the MIT project Simile[42], providing three components: a catalogue of the archived buildings, an "exhibit" of selected items of each building collection as well as the entire collection for deepened studies (linked to the "exhibit" via a keyword search). As a curation and preservation strategy FACADE recommends:

- creation of four derivate versions of a 3D model via migration:
  - original version: submitted version
  - display: easily viewable format (e.g. 3D PDF)
  - standard: full representation in preservable standard format (STEP/IFC)
  - triangulated: geometry in preservable standard format (IGES)
- (semi-)automated conversion processing of key design file formats (e.g. PDF) and common digital file formats (e.g. MS Office, JPEG)
- deposition of unrestricted software copies (with libraries and archives) instead of emulation for handicap of legal access).

The project executors draw the conclusion [42] that – since file systems of architectural firms are very inconsistent and inadequate – the provision of data in predetermined file formats is unrealistic. Organization and annotation must therefore be part of the project workflow. In order to gain complete data collections architects should be provided with guidelines what kind of material should be kept. Moreover intellectual property rights management is a challenge, architectural data being among the most difficult type of material. The FACADE project recommends acquiring a license for copies from the architect and negotiating retention periods ("embargos") for particular documents.

3.1.2 DEDICATE

Another small but important project in this field is the DEDICATE (Design's Digital Curation for Architecture) Framework in Architectural CAD Courses Design conducted

[42] Simile ran from 2003 to 2008. It was targeted on the interoperability of different digital collections, oriented towards Semantic Web technology and standards such as RDF (Resource Description Framework).

by the University of Glasgow's HATII (The Humanities Advanced Technology and Information Institute)[43]. DEDICATE, running from 2012 to August 2013, dealt with the curation of digital records with a special focus on Built Heritage. The focus of the project was to evaluate the current state of curation practises as well as policies and to locate future areas of research.

Facing experiences from curation studies that the heterogeneous CAD data is currently spread over various repositories without standardized policies – which, moreover, often lack specificity and disregard the target communities' requirements – DEDICATE followed initiatives like NINCH (Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials) and 3D-COFORM (Tools & Expertise for 3D Collection Formation) to develop policies adhering to user needs and contextual integration. As an outcome, future research questions regarded in particular:

- capture methods, modelling tools and data formats as well as information to be kept in the metadata for the ingest into a repository
- policies for an evaluation method for the choice of objects to be ingested
- ingestion processes and automated procedures
- a digital asset management architecture
- a model for interoperability (preservation of original functionalities, handling of intellectual property rights)
- transformations put to original data and management of data migration (adequate metadata, rights management).

In addition the project aimed at strengthening the interoperability and reusability of CAD data thus addressing the increasing necessity for legal and authoritative digital data management. This is a reaction to "recent and international regulations enforcing the digital documentation of public works in BIM formats"[44]. Based on audits of existing repositories conducted by its research partners and in collaboration with its tar-

[43] http://www.gla.ac.uk/colleges/arts/knowledge-exchange/themes/digital/dedicate/. DEDICATE was funded by the Arts and Humanities Research Council (AHRC) and the University of Glasgow.

[44] From http://www.gla.ac.uk/colleges/arts/knowledge-exchange/themes/digital/dedicate/

get communities (architectural practises, engineering consultancies and building control authorities) DEDICATE defined a curational workflow.

Although an important part of preservation strategies and promoted by internationally acknowledged ontologies (e.g. CIDOC-CRM)[45] , metadata annotations and semantic browsing are still considered as neglected in the area of digital Built Heritage by the DEDICATE executors. In their opinion the London Charter does indeed represent a "major advance [. . .] in the fields of descriptive and structural metadata for 3D documentation of Built Heritage but [. . .] did not affect either the accessibility or the preservation of these digital objects"[46].

3.1.3 3D-COFORM

The 3D-COFORM (Tools & Expertise for 3D Collection Formation) Large Scale Integrating Project[47] was an initiative to cover the entire processing chain connected to 3D documentation of cultural heritage artefacts for scientific and cultural purposes. Research scenarios involved:

- search and retrieval of artworks
- archaeological and historic urban site modelling
- digital reconstruction and restoration
- complicated material acquisition
- annotation

The project lasted from 2008 to 2012 and brought together a total number of 19 partners from various areas including curators, museums, and computer scientists from different areas of digitization. This enabled 3D-COFORM participants to digitize a wide range of artefacts including objects with difficult surface properties and to link the raw data to other types of information including textual metadata which were partially acquired

[45] ISO 21127:2006: Information and documentation – A reference ontology for the interchange of cultural heritage information

[46] From: `http://architecturedigitalcuration.blogspot.co.uk/p/blog-page.html` (Research Context).

[47] `http://www.3d-coform.eu` 3D-COFORM was funded within FP7 of the European Union

from other collections like Europeana[48]. For this purpose 3D-COFORM developed its own Repository Infrastructure (RI) that allows the storage of the complete digital provenance of the object. It was based on the extensible reference model CIDOC-CRM (ISO 21127:2006) to facilitate the generation of coherent metadata, temporary data management and tool monitoring [18].

Apart from the central building blocks of acquisition, analysis, and presentation, the 3D-COFORM consortium also investigated ways to ensure long-term digital preservation of the newly acquired content on the basis of the repository model. The preservation manager – partly integrated with the RI – was designed as a suite of three main preservation components [3]:

- the Preservation Information Package Manager for the conceptual composition of the archival packages
- the Preservation Risk Manager for the monitoring of risk relationships and the securing of information accessibility
- the Preservation Dependency Manager for the assignation of structural and semantic relations for the representation of objects and metadata

After the completion of the project, former members founded the Virtual Competence Centre for 3D in cultural heritage (VCC-3D)[49], a non-profit community interest company that focuses on further dissemination and exploitation of 3D-COFORM's results.

3.1.4 LOTAR

The ongoing LOTAR (LOng Term Archiving and Retrieval)[50] project brings together aerospace and defence companies from Europe and the Americas. The LOTAR international project came forth out of the IAQG (International Aeropspace Quality Group) in 2008 and follows the objective to develop, test, publish and maintain standards for the respective industries' digital data, particularly focusing on 3D CAD and PDM (product data management) data. The standards put forth by the LOTAR project are published as the EN9300 standards and as National Aerospace Standards (NAS).

[48] http://www.europeana.eu

[49] http://www.vcc-3d.com

[50] http://lotar-international.org

LOTAR bases its approach on standardized processes and models, following the OAIS reference model and the STEP application protocols AP203 (Application Protocol for Configuration Controlled Design, ISO 10303-203)[51] and AP214 (Application Protocol for Core Data for Automotive Mechanical Design Processes, ISO 10303-214)[52]. The project regards the STEP format as the "currently most advanced open format" which provides branch specific needs, enables data exchange and includes methods for specification as well as for conformance testing.[53].

The standards which are developed within LOTAR can be divided into three groups.

- "Basic Parts" which give a basic overview and outline fundamental requirements and methods[54]
- "Common Process Parts" which define the main functional entities in-line with the OAIS, i.e., ingest, archival storage, retrieval as well as e.g., data preparation[55]
- "Data Domain Specific Parts" which deal with requirements of specific information and data types[56]

Within the LOTAR project, the standardization work is currently being actively worked on in six working groups:[57]

- 3D CAD with PMI
- PDM

[51] http://www.steptools.com/support/stdev_docs/express/ap203/

[52] http://www.steptools.com/support/stdev_docs/express/ap214/index.html

[53] See http://www.lotar-international.org/lotar-organization/fundamentals-processes.html

[54] So far six "basic parts" standards have been released: prEN/NAS 9300-002: Requirements, prEN/NAS 9300-003: Fundatmentals and concepts, prEN/NAS 9300-004: Description Methods, prEN/NAS 9300-005: Authentication and Verification and prEN/NAS 9300/007: Terms and References. See http://www.lotar-international.org/lotar-standard/overview-on-parts.html-BasicParts

[55] So far six "common process parts" have been released: prEN/NAS 9300-010: Overview Data Flow, prEN/NAS 9300-011: Data Preparation, prEN/NAS 9300-012: Ingest, prEN/NAS 9300-013: Archival Storage, prEN/NAS 9300-014: Retrieval and prEN/NAS 9300-015: Removal. See http://www.lotar-international.org/lotar-standard/overview-on-parts.html-CommonProcessParts

[56] So far three "data domain specific parts" have been released: prEN/NAS 9300-100: Fundaments and concepts, prEN/NAS 9300-110: Explicit Geometry and prEN/NAS 9300-115: Explicit Assembly Structure. See http://www.lotar-international.org/lotar-standard/overview-on-parts.html-BasicParts-DataDomainSpecificParts

[57] http://www.lotar-international.org/lotar-workgroups.html

- Composites
- Electrical
- 3D Visualization
- Metadata for Archive Packages

The 3D CAD with PMI (Product and Manufacturing Information) group follows the goal of the preservation of explicit 3D geometric shape representation and associated PMI data. In the specification of processes for this objective, two main information levels were defined: the “representation level”, which include the PMI in a machine-readable STEP file intended for data exchange and the “presentation level” which includes the PMI in a form which is interpretable by the user when viewing the 3D model. The “presentation level” can be further broken down into a “polyline presentation”, which breaks down the information in lines and arcs, as well as into a “semantics presentation”, which describes the information in regards to positioning and styling.[58]

3.1.5 London Charter for the Computer-based Visualisation of Cultural Heritage

While MIT FACADE, DEDICATE and 3D-COFORM are finished and LOTAR is an ongoing project, the London Charter is a guideline which identifies good practise principles for the computer-based visualisation of cultural heritage. The idea to set up a charter of principles underlying the employment of three-dimensional visualisation technologies in heritage research and distribution and to establish them as a research method emerged from a symposium at the British Academy London in February 2006. The symposium “Making 3D Visual Research Outcomes Transparent”[59] was held in the context of EPOCH (European Network of Excellence in Open Cultural Heritage)[60]. The main principles of the London Charter were established during a subsequent seminar at King’s College London.

Originally titled The London Charter for the Use of 3D Visualisation in the Research and Communication of Cultural Heritage, it focused on the use of 3D data in an academic or

[58] http://www.lotar-international.org/lotar-workgroups/3d-cad-with-pmi.html

[59] http://www.kvl.cch.kcl.ac.uk/Symposium/index.html

[60] http://www.epoch-net.org

curatorial context in the beginning. First drafts were published in March (version 1) and June 2006 (1.1). After a meeting of the Advisory Board in Brighton in November 2007, a second draft (2) was released in February 2008, which included the renaming of the charter to its present title – The London Charter for the Computer-based Visualisation of Cultural Heritage – with its extension of scope: Since then the charter is not limited to 3D any longer, but includes all types of visualisation both 2D, 3D and 4D in the form of hard-copy printouts as well as physical objects from 3D printers (e.g., reproductions of artefacts). Beyond the academic/curatorial context it additionally targets the educational and commercial field now as well, including the entertainment sector. A revision of the charter followed in February (2.1), which is the currently valid version. Since April 2009 it has been worked on further (2.1.1).

The charter's objectives are [16]:

- to "provide a benchmark"
- to "promote intellectual and technical rigour"
- to "ensure that computer-based visualisation processes and outcomes can be properly understood and evaluated"
- to "enable computer-based visualisation authoritatively to contribute to the study, interpretation and management of cultural heritage assets"
- to "ensure access and sustainability strategies"
- to "offer a robust foundation upon which communities of practice can build detailed London Charter Implementation Guidelines".

The charter limits itself to identifying broad principles instead of giving tight regulations. As an expanding range of visualisation methods and research aims are expected in the future, those principles regard:

1. Implementation
2. Aims and methods
3. Research sources (intellectual integrity)
4. Documentation (reliability)
5. Sustainability

6. Access

Following the recommendation of the London Charter to set up specific guidelines for the execution in different subject communities, a first implementation has been started for the archaeological community: The Seville Charter is currently being drafted by the International Forum of Virtual Archaeology (set up by the Spanish Society of Virtual Archaeology SEAV)[61]. An initial draft was published in 2008. There are no other implementations in subject communities so far that the London Charter Initiative is aware of.

Current and future activities of the London Charter Initiative concentrate on the exploration of formal endorsement (e.g., by ISO, UNESCO). A second priority is the implementation in collaborative online environments: The Project LCSL (The London Charta in Second Life), funded by the British Council and the Italian Ministry for Research and Universities, looks into the necessary combination of conceptual and technological developments. Another scope is the setting up of an international online index of heritage visualisation projects, e.g., the 3DVisA Index of 3D Projects (by Anna Bentkowska-Kafel)[62]. Initiatives like, e.g., V-MUST Virtual Museum Transnational Network as part of the EU 7th framework programme advance these endeavours on the basis of the London Charter terms[63].

3.1.6 Related Virtual Heritage Projects

Related to the London Charter there are many initiatives in the growing sector of Virtual Heritage right now, such as ITN-DCH – Initial Training Networks for Digital Cultural Heritage (started in 2013)[64] or the diverse projects of the King's Visualisation Lab, King's College London[65]. Under the coordination of the Albert Ludwigs Universität Freiburg the project ROVINA – Robots for Exploration, Digital Preservation and Visualization of Archaeological Sites[66] was launched in 2013. It concentrates on the autonomous mapping and digitizing of archaeological sites, especially those that are inaccessible by humans.

[61] http://www.arqueologiavirtual.com/seav/

[62] http://3dvisa.cch.kcl.ac.uk/projectlist.html

[63] http://www.v-must.net/, see also http://www.kcl.ac.uk/artshums/depts/ddh/research/projects/current/vmtne.aspx

[64] http://www.itn-dch.eu/

[65] See: http://www.kvl.cch.kcl.ac.uk/projects.html

[66] http://www.rovina-project.eu/project

Applying robotics, accurate and textured 3D models including annotations and semantic information shall be gained. Therefore the project also aims at developing software components, opening new commercial applications of robots. Though addressing “digital preservation tools” and “digital preservation” in the project title, ROVINA seems to define “digital preservation” not in the sense of “preserving digital material”’, but in the sense of “preserving the (archaeological) site” instead.

3.1.7 Related Preservation Process Projects

Two preservation process projects with a relevance to 3D data were SHAMAN and KIM, which both focused on PLM (product lifecycle management) data.

SHAMAN (Sustaining Heritage Access through Multivalent Archiving)[67] ran between 2007 and 2011 and targeted the future accessibility of socially valuable digital objects of any kind. It delivered integrated tools to a wide mixture of target groups for the management of storage, access and presentation which were tested and validated in three application domains dealing with different types of objects. These regarded scientific publishing and government archives, industrial design and engineering (e.g., CAD) as well as e-science resources.

In regards to 3D data SHAMAN conducted a demonstration and evaluation of their framework with Philips Consumer Lifestyle division. The project evaluated Philips’ “ideation” product lifecycle management process, which involves different actors who contribute information in varying formats (e.g., .doc, .jpeg, .xls) to the product development chain. For the CAD model representation itself, SHAMAN chose the JT file format. For the Philips presentation, the SHAMAN framework used the mutlivaltent fab4browser[68] which allows the viewing and annotation of the process relevant file formats including JT. SHAMAN integrated the fab4browser into the iRODS data cloud, which formed the basis of the SHAMAN architecture framework[21]. The evaluation of the Philips case study showed that while it was clear that the framework is applicable to the domain, the particular focus group did not rate the need for digital preservation as highly as anticipated[36].

The KIM (Knowledge and Information Management through Life)[69] project was a UK

[67] http://shaman-ip.euSHAMANwas part of the EU’s Seventh Framework Programme.

[68] https://code.google.com/p/fab4browser/

[69] http://www-edc.eng.cam.ac.uk/kim/

based project which ran from 2006 to 2009 and focused on PLM (Product Lifecycle Management) in engineering domains, particularly shipbuilding, aerospace and civil engineering. A total number of 13 UK university departments were involved - 8 of which classify as "Innovative Manufacturing Research Centers" (IMRCs). Expertise on digital information management and digital curation was contributed particularly through the University of Bath, UKOLN.

The project activities were based on the notion of a shift from product approach to a service approach, where product support and service around the product has to be guaranteed for 30-50 years. Based on this notion three work packages were formulated, which covered the following aspects:

- Work package 1: capturing and recording mechanisms of information produced during creation
- Work package 2: ongoing curation of object with goal to better understand value of information
- Work package 3: organizational impacts, e.g. human resource implications of move to service approach or required decision support mechanisms

As part of this shift, a number of demands for capturing PLM information in CAD models were identified, e.g., the protection of commercially sensitive information which calls for a differentiation between public and private views of the data, the ability to generate representations from different view-points to assist different processes, a high technological interoperability level to assist the rapid sharing of information across distributed systems and platforms and also support of long-term preservation. In order to meet these requirements, the KIM project proposed capturing the PLM information in lightweight representation formats which should be annotated throughout their lifecycle. As potential lightweight CAD model representations, the file formats 3D XML, JT Format, PLM XML, PRC, Universal 3D (U3D), X3D and XGL/ZGL were explored[33][32].

The project recognized that these formats have different strengths and weaknesses. To aid a user in choosing the format best suited for the specific needs, a decision making support tool called the RRoRIfE (Registry/Repository of Representation Information for Engineering) was developed. The tool is based on a representation information characteristics ontology which has been applied to the various characteristics of file formats as well as the respective conversion software. A second prototype development within the KIM project was the annotation software LiMMA (Leightweight Models with Multi-layered Annotations). LiMMA specifically targets the problem of capturing and storing additional information by allowing users to add information layers at different lifecycle stages and link it to the lightweight representation by attaching unique identifiers to the respective entities[33][32].

3.1.8 Related Linked Data Projects

In the field of Linked Data PRELIDA and DIACHRON are notable projects. Started in 2013 PRELIDA (Preserving Linked Data) is a coordination action of the EU's Seventh Framework Programme, coordinated by the Institute of Information Science and Technologies (Consiglio Nazionale delle Ricerche – CNR, Italy)[70]. It concentrates on the discussion of existing solutions for the preservation of Linked Data and future requirements regarding quality, usability and maturity. Also targeted are specific characteristics of the Linked Data cloud in terms of structure, interlinkage, dynamicity and distribution. Bringing together end users and providers of data, services or technologies with preservation professionals – in workshops, consultations and via an online-platform – the project aims at raising awareness for preservation issues within the Linked Data community thus identifying new research questions. The greater objective is the development of a road map for the detected needs for action.

DIACHRON (Managing the Evolution and Preservation of the Data Web), coordinated by INTRASOFT INTERNATIONAL SA, has also just been set off in 2013[71]. Under the supposition that the process of publishing data is the same as the process of preserving data, DIACHRON deals with the preservation of (semi-)structured, evolving and coherent data. It engages in the creation of effective and efficient techniques for the management of the web data lifecycle and aims at the improvement of the data by temporal and provenance annotations. The automated acquisition and annotation of metadata, especially that describing provenance and all forms of contextual information is central to the project. Modules to be developed regard acquisition, annotation, evolution and archiving (including longitudinal query processing and multiversion archiving). The eval-

[70] `http://prelida.eu/` – Project partners include the Europeana Foundation and APA (European Alliance Permanent Access)

[71] `http://www.diachron-fp7.eu/` – The project will be finished in 2016.

uation of the project's outcome is intended by application to three use cases covering open governmental data lifecycles as well as large enterprise data intranets and scientific data ecosystems in the life-sciences.

3.1.9 Related Guidelines and strategies

The development of specific guidelines and strategies is another focus of various projects. The digital arts and humanities project – AHDS (Arts and Humanities Data Service)[72] – released its Guides to Good Practice for CAD in 2000 as well as the AHDS Database of ICT Projects and Methods.

Of particular importance here are the project Heritage3D (Developing professional guidance – laser scanning in archaeology and architecture) running from 2004 to 2006 and its successor, the Heritage 3D project (2008-2011)[73], both conducted by English Heritage in cooperation with the School of Civil Engineering and Geosciences at Newcastle University. The project's objectives were the support of archaeologists, local planning authorities, instrument manufacturers and software developers concerning the use of 3D laser scanning and the development and establishment of best practises in laser scanning. The allocation of impartial information on 3D survey and recording as well as on specific applications and techniques to professionals engaged in cultural heritage was a central feature, too. As delivery file formats Heritage3D suggests DXF or DWG for CAD drawings and text based grid formats for digital terrain models; it also provides a minimum set of descriptive metadata for raw point cloud scan data[74]. An extended set of descriptive metadata based on English Heritage's specifications is given in Laser Scanning for Archaeology, A Guide to Good Practice[75] by the ADS.

Heritage3D also refers to the guidelines assembled by the Archaeology Data Service's (ADS) project Preservation and Management Strategies for Exceptionally Large Data Formats, commonly known as the "Big Data project" (in cooperation with English Heritage). In the project's final report (2007)[76] "big data" is defined as the growing size

[72] http://www.arts-humanities.net/

[73] http://www.heritage3d.org/. The projects were funded by the National Heritage Protection Commissions programme (formerly: Historic Environment Enabling Programme).

[74] http://www.english-heritage.org.uk/publications/3d-laser-scanning-heritage2/, see also: Andrews (et al.): Metric Survey Specifications for Cultural Heritage, 2009.

[75] http://guides.archaeologydataservice.ac.uk/g2gp/LaserScan_Toc

[76] http://archaeologydataservice.ac.uk/attach/bigData/bigdata_final_report_1.3.pdf

of data sets (giga- and terabytes) created by archaeologists through technologies such as Lidar (Light Detection and Ranging or Laser Imaging Detection and Ranging), 3D laser scanning, maritime survey (sidescan sonar, sub bottom profiling and others) and digital video. An "information object" is specified as comprising its "content data object" as well as its "representation information". The "Big Data project" also provided a sample of formats considered employable for long term preservation.

3.2 Bit Preservation

While processes ensuring the safety of bits have been implemented in good IT practise for quite a while, bit preservation cannot be automatically considered solved. As Strodl et al. point out, failure of hardware and storage media as well as human error are inevitable and will remain an endangerment to long-term bit integrity [43]. While technological means such as integrity checking and redundant storage are available, the responsibility of implementing and controlling the mechanisms are on the organisational side of digital preservation. There, requirements for bit storage, e.g., transmissions speed or operation procedures, need to be carefully evaluated and chosen[49]. While this is true for all data, no particularities exist for the bit preservation of 3D data which do not hold true for any form of data.

3.3 Logical Preservation

The MIT FACADE project noted that it was difficult to obtain information about CAD file formats internal characteristics, as the CAD software providers were not willing to publically release this information for obvious commercial reasons [42]. However, this information is necessary to develop tools and mechanisms for file format identification, characterization and validation. The DURAARK project foregoes this by concentrating on existing open file formats: IFC-SPF and E57. While IFC-SPF covers the "as-planned" data produced in CAD software, E57 is a file format documenting the "as-is" state of objects through 3D scanning procedures. Both file formats are open standards, whose sustainability factors are further analyzed in chapter 3.3.2. MIT FACADE [42] as well as the recent DPC (Digital Preservation Coalition) technology watch report "Preserving Computer-Aided Design" recommend open file formats, such as IFC (Industry Foundation

Classes) or STEP (STandard for the Exchange of Product Data), for archiving the full model information [7].

The following section will include a detailed analysis of the file format sustainability factors and briefly touch on the availability of tools for the logical preservation of IFC-SPF and E57 files.

3.3.1 File Format Identification

File format identification can be split into two requirements: the file format should be registered in one of the file format registries maintained by the digital preservation community (PRONOM[77], UDFR[78]) and a tool should exist which can identify the format on a granularity level of the format's version, if applicable. Within the MIT FACADE project a few native CAD file formats, such as file formats associated with AutoCAD 2004-2005 and 2007-2008, CATIA 4 and CATIA 5 as well as Revit and SketchUp were submitted to the TNA (The National Archive, UK) to be included in the PRONOM file format database [42]. As the semantically enabled Unified Digital Format Registry (UDFR) imports PRONOM information, the native CAD file formats are registered there as well. Neither file format registries, however, have entries for IFC or E57.

The DURAARK project ran a small test set of e57 files collected from the libE57 site [79] and IFC files collected from the IFCWiki [80] against identification tools.

The test set ran through file format identification using two tool sets: Fido version 1.0.0 [81] and Fits version 0.6.2 [82]. Fits is a framework which wraps multiple digital preservation file format characterization tools and normalizes the output. The version used in the test included Jhove, Exiftool, the NLNZ Metadata Extractor, DROID, FFIdent and the File utility for the test. The DROID signature pattern was updated to version v72.

As an outcome, fido reported "fail" for all test data - neither E57 nor IFC-SPF could be identified. Fits reported the e57 files as "unknown binary" failure status. The IFC-SPF

[77] http://www.nationalarchives.gov.uk/PRONOM

[78] http://www.udfr.org

[79] http://www.libe57.org/data.html

[80] http://www.ifcwiki.org/index.php/Examples

[81] https://github.com/openplanets/fido

[82] http://code.google.com/p/fits/

files were recognized as plain text/ASCII files, which is correct on the encoding level but not on the file format level.

3.3.2 File Format Sustainability

In the following section, the file formats E57 and IFC-SPF will be described against the file format sustainability factors described in chapter 2.3.1. While IFC-SPF covers the "as-planned" data produced in CAD software, E57 is a file format documenting the "as-is" state of objects through 3D scanning procedures.

E57

1. Disclosure

   The E57 file format specification is available as the ASTM E2807–11 standard. It is developed to be a well-documented, open and vendor-neutral standard.

   - *well documented and complete specification:*
     The specification contains a concise and complete description of the file format as well as necessary mathematical definitions.
   - *public (open) specification:*
     The specification is available for purchase at the ASTM website [83].
   - *format specification should be stable and - if changes occur - backward compatible:*
     The current version 1.0 has been stable since February 2011.

2. Internal technical characteristics

   The basis of the E57 file format is an extensible XML structure for storing metadata of one or multiple point clouds and associated data (e.g., images taken during the scanning process) within a single file. Data parts of an E57 file are stored in binary formats which are also part of the specification.

   - *free from encryption:*
     The E57 format is free from encryption.

[83] http://www.astm.org/Standards/E2807.htm

- *free from Digital Rights Management (DRM) copy protection:*
  The E57 format is free from DRM (Digital Rights Management) copy protection.
- *complexity should meet the intended functionality and not be over-specified:*
  The file format's complexity meets the requirements for efficiently storing large amounts of data as well as associated metadata. In addition, the format offers an extension mechanism to allow customizations of the format. To achieve this, the format is a combination of binary data and XML (eXtensible Markup Language).
- *error-detection included in format:*
  The format uses CRC32C checksums throughout the physical file to maintain data integrity.

3. External technical characteristics

   The E57 file format does not make specific assumptions on the hardware, software or storage medium used. It may be used on any relevant computing platform.

   - *independent of hardware:*
     The reference implementation libE57[84] is implemented using the C++ programming language which enables its use on virtually any relevant platform.
   - *independent of physical medium:*
     Files in E57 format may be stored on any relevant kind of medium (e.g., local files, remote files, database).
   - *independent of specific software or operational system:*
     The file format may be parsed on any relevant operating system (e.g., Linux, Windows).
   - *independent of external information:*
     The file format is independent of external information. The only information referenced via URI in the XML section of the file format is the namespace.

4. Format Acceptance

   The E57 file format is being adopted by an increasing number of software vendors

[84] http://libe57.org/

(Autodesk Inc., Bentley Systems Inc., FARO Technologies Inc., Zoller+Fröhlich GmbH, among others) for inclusion in their respective software packages.

- *support through several software manufacturers:*
  The E57 format is supported by several software vendors, a list of current partners is available on the libE57 website[85].
- *embraced / popular with industry:*
  Even though the file format is relatively new, many popular software packages have already integrated support for E57 files[86].
- *used by several domains:*
  Being a file format for point cloud data, usage of the file format depends on the software which supports the format. Among the software which currently supports E57 are tools for architecture, construction, cultural heritage, forensics, and other tasks.
- *standardised (ISO, SIS, etc.):*
  The format is standardized as ASTM E2807–11.

5. Patent

   The E57 standard was developed to be an open and vendor-neutral standard for storing point cloud data. An open-source reference implementation (libE57) is freely available.

   - *free from patent / licensing costs:*
     Usage of the E57 format is free from patent or licensing costs. The libE57 implementation of the standard is open source[87].

6. Logical Structure and Transparency

   The E57 file format is well-defined and – given its versatility – can be understood and parsed in an adequately simple manner using the specification and available software libraries.

   - *existing methods for validation of file structure:*

[85] http://libe57.org/partners.html
[86] http://libe57.org/products.html
[87] http://libe57.org/license.html

The libE57 implementation of the standard includes a file validation tool (e57validate).

- *self-documented format, containing i.e., descriptive metadata:*
  The XML portion of E57 files which stores scan metadata uses descriptive field and section names which make this part of the format mostly self-documenting. Documentation of the binary parts within the file is not part of the specification.
- *the file's content is transparent for "simple" tools:*
  For meaningful access to all data included in an E57 file, a software library like libE57 should be used. The library also includes tools for extracting parts of a given E57 file (e.g., the XML part, individual fields of metadata, point data, image data).
- *standard or simple representation of the data in the file (e.g., human readability):*
  The XML part of an E57 file is human-readable. A tool for splitting the XML and binary parts automatically is part of libE57.

IFC-SPF

It needs to be noted that three IFC file format variants exist: IFC-SPF, the STEP physical file data encoding of an IFC file as defined by ISO 10303-21; IFC-XML the eXtensible Markup Language data encoding of an IFC file as defined by ISO 10303-28 and IFC-ZIP, a PKzip 2.04g compressed version of either an IFC-SPF or IFC-XML encoding[88]. As IFC-SPF is the STEP Part 21 version of the IFC file format and the one chosen for the DURAARK scope, the sustainability factors will only regard IFC-SPF.

1. Disclosure

   The IFC-SPF format and schema specification is available through the website of the buildingSMART Foundation.[89]

   - *well documented and complete specification:*
     The specification contains a concise and complete description of the schema.

[88] see http://www.buildingsmart-tech.org/specifications/ifc-overview/ifc-overview-summary

[89] http://www.buildingsmart.com

  Also, examples that implement the specification are included in the annex of the document.

  - *public (open) specification:*
    The specification is openly available at the buildingSMART association's website. [90]

  - *format specification should be stable and - if changes occur - backward compatible:*
    The releases IFC1.0, IFC1.5.1 and IFC2.0 were in use between 1996 and 2000 and are now considered outdated and no longer supported. IFC2x (October 2000) was superseded by IFC4 in 2013. However, IFC2x remains supported by IFC4 Tools. The IFC4 specification includes a change log which compares the IFC4 specification to the previous version IFC2x3 TC1 and documents the elements that were added, modified or deleted.

2. Internal technical characteristics

   IFC-SPF is encoded as a structured ASCII text file, where the data within the text is structured in the EXPRESS information modelling language[91]. EXPRESS is implementation-independent and standardized in ISO 10303-21.

   - *free from encryption:*
     The IFC-SPF format is free from encryption.

   - *free from Digital Rights Management (DRM) copy protection:*
     The IFC-SPF format is free from DRM copy protection.

   - *complexity should meet the intended functionality and not be over-specified:*
     As the purpose of the file format is to support various participants in a building construction or facility management project, the schema is extensive and overly complex. To manage this high degree of complexity, IFC-SPF supports Model View Definitions (MVD). A MVD allows a subset view of the data model and supports one or more domain recognized workflows. BuildingSMART makes official model views available on their website[92] and supplies further

[90] `http://www.buildingsmart-tech.org/downloads/ifc` (registration necessary)
[91] `http://www.buildingsmart-tech.org/implementation/faq/fag-general-ifc-spec`
[92] `http://www.buildingSMART-tech.org`

information on related specifications. Additionally, users can define their own MVDs.

- *error-detection included in format:*
  The error-detection of the format itself is limited to syntax descriptions in entity definitions. Here "where" rules are included to guarantee the correct use of data structures, for instance in the case of data type restrictions.

3. External technical characteristics

   The IFC file format does not make specific assumptions on the hardware, software or storage medium used. It may be used on any relevant computing platform.

   - *independent of hardware:*
     IFC-SPF has no hardware dependency.
   - *independent of physical medium:*
     IFC-SPF is not bound to a specific data carrier.
   - *independent of specific software or operational system:*
     As the primary use for IFC-SPC is data exchange, it is not bound to a specific software or operational system
   - *independent of external information:*
     IFC-SPF may reference external information via a uniform resource identifier (URI) such as a uniform resource name (URN) or uniform resource locator (URL).

4. Format Acceptance

   IFC-SPF is a STEP family file format. STEP is the informal notation for the international standard for the computer-interpretable representation and exchange of product model data: the ISO-10303 family of standards. Different domains have adopted STEP in domain-specific implementations, such as the IFC formats. The main current usage of IFC-SPF is as a data exchange format. Out of the IFC formats, IFC-SPF is the most widely adopted format.

   - *support through several software manufacturers:*
     An extensive list of AEC CAD software vendors supports IFC-SPF as an

import and/or export format[93]. Furthermore, there are several viewers and converters[94] as well as a few open source tools[95] available.

- *embraced / popular with industry:*
  A number of commercial AEC CAD software packages support IFC-SPF as an import and/or export format. Furthermore, buildingSMART offers a certification process for Import and Export routines of software applications. The list of certified applications is available through the buildingSMART webpage[96].
- *used by several domains:*
  While the IFC formats cater only to the AEC domain, the STEP basis of the format is widely adopted with application protocols for several domains, such as the steel construction industry (CIMsteel) or the aerospace industry.
- *standardised (ISO, SIS, etc.):*
  The format is a STEP file format / application protocol. STEP is standardized in ISO 10303, IFC makes use of the standard parts STEP-Part 11 (EXPRESS Language reference manual), STEP-Part 21 (STEP-File) and partly STEP-Part 42 (Geometric and topological representation). The IFC format is standardized as ISO 1739:2013:“Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries”.

5. Patent

   The IFC standard was developed to be an open and vendor-neutral standard to be used as a data import and export format for a variety of CAD software.

   - *free from patent / licensing costs:*
     Usage of the IFC-SPF format is free from patent or licensing costs. A model implementation guide is available at the buildingSMART website[97].

6. Logical Structure and Transparency

[93] see http://www.buildingsmart-tech.org/implementation/implementations

[94] see http://www.ifcwiki.org/index.php/Freeware

[95] see http://www.ifcwiki.org/index.php/Open_Source

[96] ttp://www.buildingsmart-tech.org/certification/ifc-certification-2.0/ifc2x3-cv-v2.0-certification/participants

[97] http://www.buildingsmart-tech.org/implementation/ifc-implementation/ifc-impl-guide/ifc-impl-guide-summary

The IFC-SPF format is well-defined through its available schema. As a clear text format IFC-SPF is human readable and transparent to methods for validation.

- *existing methods for validation of file structure:*
  The buildingSMART "Global Testing and Documentation Server" (GTDS)'[98] offers a validation of the IFC-SPF file structure.
- *self-documented format, containing i.e., descriptive metadata:*
  IFC-SPF is self-documented through its schema.
- *the file's content is transparent for "simple" tools:*
  As a structured ASCII text file the IFC-SPF file content is transparent for "simple" tools.
- *standard or simple representation of the data in the file (e.g., human readability):*
  As a structured ASCII text file, the content of the IFC-SPF file is human readable.

3.3.3 Technical Metadata Extraction

The MIT FACADE project mentioned that Jhove was included as a technical metadata extractor within the workflow [42], however, as Jhove does not include modules for native CAD formats or 3D exchange formats, technical metadata extraction was only conducted for standard file formats such as PDF within the project. As part of the small experiment conducted in the DURAARK project and described in section 3.3.1, technical metadata extraction through the tools wrapped in the fits framework was also tested.

Since DROID identified IFC-SPF as an ASCII file, the Jhove instance wrapped in fits consequently used the ASCII module to extract technical metadata from the IFC files. As a result, text/plain was reported as MIME type with charset=US-ASCII. The ASCII module also supports the extraction of line ending and additional control characters. For the IFC samples, CRLF was extracted as a line ending. No additional control characters were found. No technical metadata was extracted for E57.

For E57 files the open source reference implementation libE57 [99] includes two tools which

[98] http://gtds.buildingsmart.com/

[99] http://www.libe57.org

may allow capturing of technical metadata: e57fields and e57xmldump. Each E57 file contains an XML section which describes the hierarchy of the file as well as some basic values - the e57xmldump tool allows the extraction of the entire XML section. The e57fields tool allows to generate statistics about the files fields usage, indicating fields count and min as well as max values.

3.3.4 Validation

As mentioned in the previous sections, Jhove identifies IFC-SPF files as ASCII files. All IFC files in the sample set were considered well-formed and valid ASCII files. Based on the Jhove ASCII module description[100] conformity to well-formedness and validity of ASCII files is fulfilled when the file consists entirely of properly ASCII-encoded text by ISO/IEC 646, ANSI X3.4, ECMA-6 specification. However, while it is helpful to know that the IFC file is encoded as a valid ASCII file, it says nothing about the IFC validity itself. Rather, ASCII well-formedness and validity should be a pre-requisite for IFC well-formedness and validity.

BuildingSMART itself offers a validation service of IFC-SPF file structures through the “Global Testing and Documentation Server” (GTDS)[101]. The GTDS server includes the ifcCheckingTool for validation, which is developed by KIT Karlsruhe Institute of Technology and also available as a “lite” stand-alone version. Well-formedness and validity of IFC-SPF files are a difficult topic. The ifcCheckingTool currently checks against Coordination View Version 1.0 which was developed between 2005 and 2009. A number of older files in the test set, e.g. an IFC file generated with ArchiCAD 7.00 could not be validated and actually caused the validator to crash.

In regards to E57 the aforementioned libE57 includes the validation tool e57validate.exe. The validator checks currently against 50 errors, some of which are basic file handling methods such as the failure to open, close or read the file. Other errors checked include the failure to represent values in the requested type, bad codecs used in CompressedVectorNode or element values being out of min/max bounds. Furthermore, the validator checks the data integrity by recalculating CRC32C checksums throughout the physical file and comparing them with the previously stored values within the file. The standard

[100] http://jhove.sourceforge.net/using.html
[101] http://gtds.buildingsmart.com/

output reports the number of errors, warning and suspicious hits per file. The informational count is always at least 0, as operation success is reported through it's own exception identifier.

Example for standard success output of E57validate.exe:

```
Informational 4000 in /e57LibraryVersion:  library version of writer:
InteliSum-LD3-Studio-V5.1-E57RefImpl-1.0.154-x86-windows
Error count:  0
Warning count:  0
Suspicious count:  0
Informational count:  1
```

Example for error output:

```
Informational 4000 in /e57LibraryVersion:  library version of writer:
InteliSum-LD3-Studio-V5.1-E57RefImpl-1.0.154-x86-windows
Error 1000 in /data3D/0/points/50/cartesianY: value 0.040011 is out of Cartesian bounds
Error 1000 in /data3D/0/points/8032/cartesianZ: value -0.061873 is out of Cartesian bounds
Error 1000 in /data3D/0/points/8323/cartesianY: value 0.187321 is out of Cartesian bounds
Error 1000 in /data3D/0/points/18586/cartesianX: value 0.061009 is out of Cartesian bounds
Error count:  4
Warning count:  0
Suspicious count:  0
Informational count:  1
```

3.4 Semantic Preservation

As described in 3.1.7 the KIM project identified the need for an annotation layer in PLM CAD models. For the architecture domain, this annotation level can already be found in semantically rich, interoperable building information models (BIM) which contain explicitly modelled 3D geometry as well as information captured during various stages of the planning and construction process. This information ranges from simple provenance information like the author to highly specific information about a particular material used in the construction process. The information in a BIM object may be self-contained but may also be contained in external datasets, such as product information datasets or classification systems provided by, e.g., vendors of specific building parts[17].

Even though a number of vocabularies, e.g., BauDataWeb[102], exist to support a structured approach to data capturing within a BIM, a great deal of information is currently modelled in a formally weak and ad hoc manner. The most notable example is the buildingSMART Data Dictionary (bsDD)[103], a reference library allowing the creation and support of multilingual dictionaries. The bsDD already contains several tens of thousands of concepts but has not been adopted widely due to limited exposure via standard interfaces. Functionality and limitations of current practise of semantic enrichment of BIM models is further discussed in deliverable D3.3.1. While the described process holds true for “as-planned” IFC data, no comparable process exists for scanned 3D data.

As mentioned in 2.4 efforts in semantically enriching archives with social web data is a new research area with projects like ARCOMEM[104] only just starting. In regards to architectural 3D models, exploitation of LOD beyond datasets of the architecture and engineering domains has not yet been exploited.

Section 2.4 further described that different sources of data follow varying patterns of evolution and exhibit different frequencies of change. For instance, intuitively the geographical coordinates of a built structure will remain largely constant throughout its life-cycle (barring rare exceptions where structures might be demolished to be reconstructed elsewhere). This would mean that apart from updates through additions to the data source (for example GeoNames), there may be little evolution with respect to the location of a structure. A street address may differ from this, where street names may change over time and the changes are relevant semantic information in interpreting the adress correctly. An even less stable concept is that of the perception of a structure. The sentiments of people and their perception of a structure are influenced by many factors and therefore exhibit sporadic evolution. Such data will need to be recomputed temporally in order to reflect constant correctness. As mentioned in 2.4 the stability of respective data sources is an area that demands further investigation.

[102] `http://semantic.eurobau.com/`

[103] `http://www.buildingsmart.org/standards/ifd`

[104] `http://www.arcomem.eu`

3.5 Metadata

The majority of domain specific metadata developments for 3D data are in the area of descriptive metadata. The MIT FACADE project developed an ontology to describe the main intellectual entity “Project Information Model” (PIM). The PIM contains an entire architectural project including not only plans and models, but also textual documentation, emails, videos or still images. The ontology focuses on the descriptive metadata level, borrowing elements from Dublin Core, where possible [42].

In the cultural heritage sector, a general framework for the description of cultural heritage resources is the International Council of Museums’ Conceptual Reference Model CIDOC-CRM [105]. It contains a proposal of a high level metadata set, which can be used for various cultural heritage objects, ranging from a bronze statue [106] to a photograph of a figure of public interest [107]. The 3DCOFORM project integrated the CIDOC-CRM core, however, metadata was mainly used on a descriptive level and to describe the relation of objects to each other [4].

English Heritage, the Historic Buildings and Monuments Commission for England, recommends the following elements to be included in a mandatory minimal metadata set for 3D laser scans of historic monuments [23]:

- file name of raw data
- scan number (unique scan number for this survey)
- date of capture
- scanning system used, including manufacturer’s serial number
- company name
- monument name
- total number of points
- point density on the object (with reference range)

[105] `http://www.cidoc-crm.org/`

[106] see example of CIDOC Core description for Rodin’s bronze statue “Monument to Balzac” `http://www.cidoc-crm.org/crm_core/core_examples/balzac.html`

[107] see example of CIDOC Core description for photograph of David Beckham at Euro 2004 `http://www.cidoc-crm.org/crm_core/core_examples/beckham.htm`

- for outdoor scanning: weather conditions during scanning

"Total number of points" and "point density on the object" are technical metadata, while the rest are descriptive metadata.

An extensive descriptive metadata schema specifically for 3D objects is the PROBADO3D[108] metadata core. The current version 1.1 contains 23 main elements, which may contain further attributes or children. It allows for the inclusion of descriptive information like creator, location, subject areas, license type (CC license types), various dates (e.g., date created, date available, date issued) and relations (e.g., isSupplementTo, IsNewVersionOf, IsPartOf). However, in regards to technical information about the object, only software name and version, file size and key appellation (units, height coordinate, vertices, polygons) are captured [9].

While some projects and organizations have started to make recommendations on technical information which should be captured in metadata, no technical metadata standard for 3D data exists. Recommendations remain on the level of one or two elements - such as in the English Heritage example above or in the case of The UK Archaeology Data Service (ADS), who recommends to flag the parts of a 3D object that are based on hypothesis rather than on evidence [28].

The MACE project (Metadata for Architectural Contents in Europe)[109] focused mainly on the connection of data from different repositories and domains for retrieval. Though no means to generate the data were discussed, the project did explore possible categories for technical metadata for geometric 3D models. These were grouped into 4 main categories:

- *creation information*
  e.g., applcation, application version
- *file format information*
  e.g., format, version, extension, size
- *geometric information*
  e.g., polygonal mesh, NURBS, parametric objects, complexity
- *file content*
  e.g., textures, shades, physically correct materials, layers, external files, lights

[108] http://www.probado.de/en_3d.html
[109] http://mace-project.eu/

Other possible categories mentioned were animation information and size and scaling information [10].

3.6 Organizational Roles in Digital Preservation

The projects and initiatives mentioned in the first section of this chapter showed that there has been a slow but steadily growing interest in the preservation of 3D data - both scans and CAD created. While permanent collaborations like LOTAR and the ADS (UK Archaeology Data Service) have taken a foothold in the engineering and archaeology domains, the architectural domain is lacking such a central collaboration platform.

Furthermore, the only known organizational implementation of digital preservation targeting 3D architectural data remains the MIT FACADE project. The main stakeholders addressed in the MIT FACADE project were those of the academic field: architecture instructors and students, with historians as a second focus group. The project considered including architecture practises as stakeholders, but stated that at the time (2006-2008) the technical environments typically used in architecture practises were not advanced enough to include digital preservation. Guidelines for archival handover from architecture practises to a long-term archive were mentioned as being extremely beneficial in the MIT FACADE file report [42], but were unfortunately not within the scope of the project and haven't been developed since.

3.6.1 Lifecycle Model

Chapter 2.6.1 named a few domain specific lifecycle models that exist. For the stakeholder definition and requirement analysis, which is described in deliverable D2.2.1, not a digital object lifecycle but one describing a typical building lifespan was used. The model allowed the definition of the different stages at which an analogue object is planned and created, changes ownership and is re-used. While it represents an analogue lifecycle it allows us to understand where data is created or changes ownership - i.e., in moving from the design phase to the pre-construction phase, where different authors are to be included. The model furthermore proved helpful in the communication with stakeholders, as it underlines the long-term aspect which needs to be taken into consideration when wanting to re-use the original data: while the "construction phase" usually lasts between 2-5 years, the “use phase” of a building has a typical length of 60 and more years. Figure 8 shows an extension of the model described in D2.2.1. It has been extended to reflect the stages at which the two 3D information objects handled in the DURAARK project are typically created and modified. While the BIM objects may be actively used throughout the entire lifecycle, ranging from the construction phase to active use during the maintenance stage, 3D scans are consequently created and used after the object has been constructed.

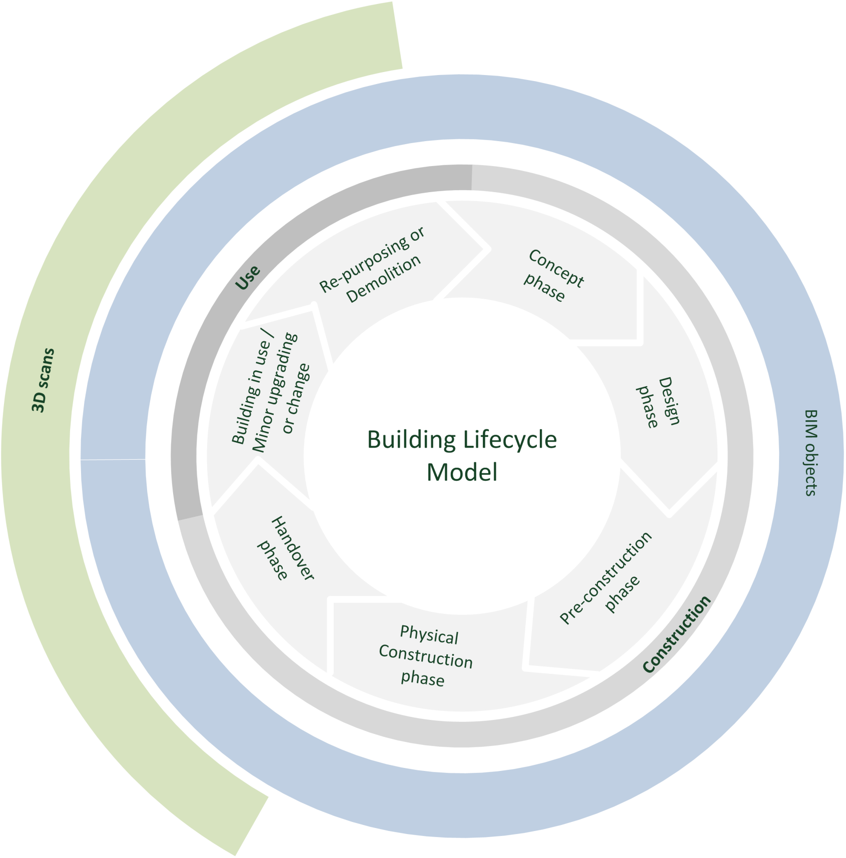


Figure 8: Building Lifecycle Model

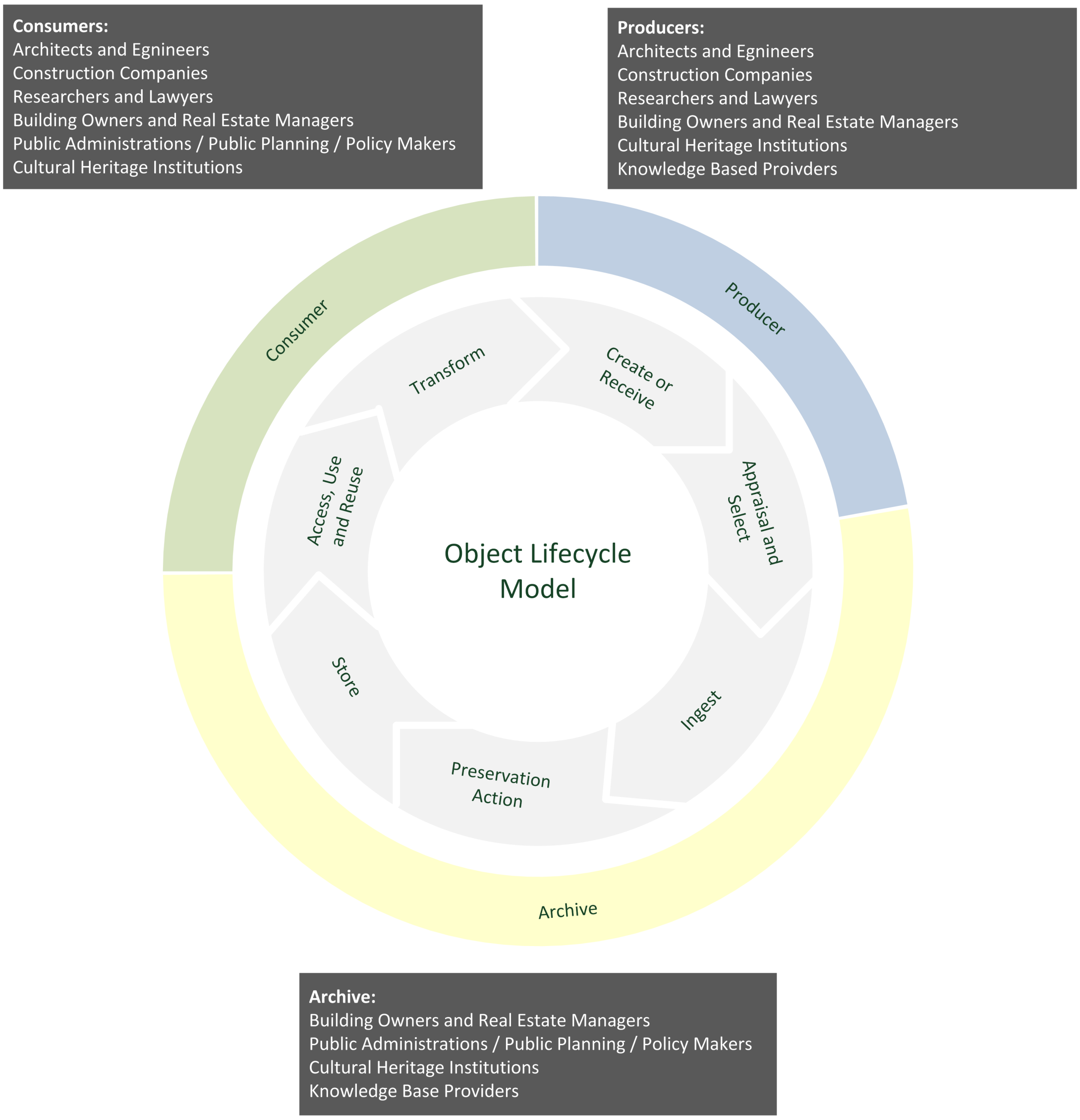
While the building lifecycle model allowed the DURAARK project to define who is typically involved in the data creation und use / re-use processes, this information was used to create a second, digital object focused model. This model, shown in figure 9 is based on the DCC curation lifecycle model described in chapter 2.6.1. It has been amended with stakeholder mappings to the roles “producer”, “archive” and “consumer”. It serves as a basic definition of the designated community in the DURAARK context and an understanding of the knowledge and context in which the digital objects were created.

3.6.2 Significant Characteristics

Chapter 2.6.2 discussed that significant characteristics and the respective significant properties may stem from three different classes: the digital object itself, the environment and the preservation action. In order to be able to describe a list of possible characteristics, a thorough analysis of those three classes is necessary. The lack of significant characteristics at the 3D digital object level is closely connected to the lack of a technical metadata standard for 3D data, as pointed out in chapter 3.5. In order to understand the environment a good knowledge of the stakeholders described above is necessary to define requirements and constraints.

3.6.3 Preservation Planning

As described in chapter 2.6.3 preservation planning requires monitoring of the community, monitoring of relevant technology and of the archive itself. The community is defined by the stakeholders mentioned in figure 8. The technology to be monitored should include the logical layer - i.e., file formats - as well as the bit layer - i.e., storage technology. Out of the defined stakeholders cultural heritage institutions form the only group which has been significantly active in digital preservation activities and therefore little is known of requirements and constraints of the other stakeholder groups.



Consumers:
Architects and Egnineers
Construction Companies
Researchers and Lawyers
Building Owners and Real Estate Managers
Public Administrations / Public Planning / Policy Makers
Cultural Heritage Institutions

Producers:
Architects and Egnineers
Construction Companies
Researchers and Lawyers
Building Owners and Real Estate Managers
Cultural Heritage Institutions
Knowledge Based Proivders

Archive:
Building Owners and Real Estate Managers
Public Administrations / Public Planning / Policy Makers
Cultural Heritage Institutions
Knowledge Base Providers

Figure 9: Curation Lifecycle Model

4 Identified gaps for the preservation of 3D objects

While chapter 2 described the de-facto standards, recommended guidelines and established best practise processes for digital preservation in a content agnostic approach, chapter 3 described the existing tools, standards and recommended guidelines for 3D data with a focus on the architectural domain and on the DURAARK relevant file formats IFC-SPF and E57. Particularly section 3.1, but also the other sub-chapters showed that digital preservation efforts targeting 3D data in general and architectural data specifically are still scarce. The following sections will list the gaps by comparing chapters 2 and 3.

4.1 Bit Preservation

Bit preservation is the first step of any preservation strategy. As described in chapters 2.2 and 3.2 bit preservation actions are part of good information technology practise. On a technological and procedural level they are well addressed through practises such as media replication, media refreshing, monitoring and fixity checks. The need for implementation of such practises on an organisational level is self-evident.

4.2 Logical Preservation

The logical layer of an object is the format level of the digital object. Preservation processes on the logical layer are therefore related to maintaining the renderability of the file. As described in chapter 2.3 standard processes of logical preservation are file format identification, technical metadata extraction and file format validation. Chapter 3.3.1 showed that IFC-SPF and E57 are neither included in file format registries of the digital preservation community, nor can they be recognized through standard digital preservation identification tools like DROID or fido. File format identification could be achieved through manual verification (e.g., IFC-SPF header), test rendering or for E57 through LibE57 tools like the e57validate.exe. However, as it is safe to assume that IFC-SPF and E57 files are to be stored in an archive which maintains various file formats, a per-format solution - especially for file identification - is not a practicable one.

The lack of technical metadata extraction tools is directly tied to the lack of a common recommendation for technical metadata for 3D data (see chapter 3.5). For E57 two potential tools for extraction exist, as mentioned in 3.3.3. The existing tools should be evaluated further for their suitability as technical metadata extractors when a clear idea of needed technical metadata exists. As described in chapter 3.3.4 the IFC-SPF validation tool ifcCheckingTool only worked for a subset of IFC-SPF files. No documentation was available. General-purpose STEP toolkits such as jSDAI[110] or the EDMdeveloper SDK[111] should be analyzed towards their suitability for logical preservation tasks such as file format validation. Regarding E57, the libE57 E57validate tool worked well for the validation of the test set. It should be pointed out, however, that the documentation for the tool is somewhat limited. While it does seem suitable based on the tests conducted, it should be analyzed further. While the sustainability factors for IFC-SPF and E57 document them as open and transparent formats which are suitable for long-term archiving, potential risks like dependency on external resources should be explored further and documented.

1. IFC-SPF is presently not described in file format registries such as PRONOM or GDFR
2. E57 is presently not described in file format registries such as PRONOM and GDFR
3. IFC formats can presently not be identified by standard file format identification tools used in digital preservation practise
4. E57 can presently not be identified by standard file format identification tools used in digital preservation practise
5. No metadata extraction tool supporting IFC-SPF available
6. Extraction tools in libE57 reference implementation (E57fields.exe, E57xmldump.exe) need to be analyzed further in regards to their suitability for techincal metadata extraction
7. Suitability of existing IFC-SPF validator needs to be evaluated further
8. E57 validator should be evaluated further and validation conditions documented in relation to standard

[110] http://www.jsdai.net

[111] http://www.epmtech.jotne.com/products/express-data-manager-edm/

9. For IFC and E57 alike, risks like dependency on external resources should be explored further and documented

4.3 Semantic Preservation

As described in chapters 2.4 and 3.1 semantic preservation is slowly being picked up in research projects. Semantic preservation research projects so far predominantly involve semantic technology experts, as in the case of DIACHRON. PRELIDA, which just started in 2013, is one of the first projects actively trying to combine digital preservation and the semantic web communities. PRELIDA's goals are to raise awareness amongst the linked data community of already existing digital preservation processes and to draft a roadmap for further linked data preservation research needs in the preservation community.

Strodl et al. point out that use of semantic technologies in general but also particularly in combination with an interdisciplinary approach are future research fields of digital preservation projects [43]. Alas, the inclusion of a content provider alongside digital preservation and semantic technology experts allows for a concrete usage scenario to further explore the so far mainly theoretical approaches of semantic preservation.

While semantic preservation efforts are currently mainly targeting the question of semantic enrichment and the preservation of the datasets themselves, the question of how to trace changing concepts which have been captured in enriched datasets has not really been addressed.

1. Relevant knowledge base sources for 3D architectural data have not been identified yet
2. Methods to calculate impact which changes of one entity within a chosen dataset graph has on other entities have not been tested in a domain specific setting
3. Methods to monitor changes within chosen datasets have not been made available in tools exposed to the domain
4. No standard method to preserve and link changing semantic concepts in metadata is available

4.4 Metadata

Chapter 2.5 showed that a basic framework of how metadata is captured has been established within the archival community. The long-term understandability of data necessarily includes the understandability of metadata. It is therefore of utmost archival interest to capture metadata in available standards wherever possible. As far as descriptive metadata is concerned, various recommendations exist ranging from slim examples, like the object section of the MIT FACADE PIM ontology, to rather extensive descriptions, as in the case of the PROBADO3D[112] metadata core. The PROBADO3D metadata core is based on Dublin Core[113] but not described using Dublin Core [9]. In order to evaluate the suitability of an existing descriptive metadata schema for the DURAARK context, the required elements at a descriptive level need to be defined. For preservation metadata, PREMIS should be used. Furthermore it is advisable to include a wrapper format which ties the information in an intellectual entity together. Here, the use of METS is recommended and should be evaluated. There is currently no registered profile specifically for 3D data available in the METS implementation registry[114]. While descriptive, administrative and preservation metadata are largely content-agnostic as described in 4, technical metadata is content specific. Currently no technical metadata schema for 3D content is available. As mentioned in 3.5 few recommendations exist, which need to be explored further. An unanswered question in this regard is whether a singular technical metadata description can cover both, IFC-SPF and E57, or whether the content types substantially differ, resulting in the need to be covered in separate technical metadata descriptions. Another topic is the question of describing semantic concepts captured as part of semantic enrichment and preservation. The suitability of existing standards needs to be evaluated.

1. Requirements in descriptive elements need to be defined and compared to existing descriptive metadata schemas
2. No de-facto standard for technical metadata of 3D data is available
3. Evaluation of whether 3D point clouds (E57) and BIM (IFC-SPF) can be described using a common group of technical metadata elements is needed

[112] http://www.probado.de/en_3d.html
[113] http://dublincore.org/
[114] http://www.loc.gov/standards/mets/mets-registry.html

4. Currently no registration of a METS implementation for 3D objects is available in the implementation registry
5. Suitability of existing standards to describe semantic concepts as part of semantic preservation need to be evaluated

4.5 Organizational Roles in Digital Preservation

Numerous sources have pointed out not only the benefit, but also the necessity of creators, users and archives working hand in hand in successful digital preservation practise [20] [47] [14] [5]. Furthermore, Strodl et al. have identified a slow shift from addressing problems within the digital preservation system on the data custodian's side towards trying to raise the creator's awareness of problems and preventing problems in their full complexity early on [43].
Addressing the problems of architectural 3D data in BIM and point cloud objects therefore requires a good knowledge of the creation, the preservation, the use/re-use processes as well as of the requirements and constraints that stakeholders have in the object.

The DURAARK project identified the stakeholders as shown in figure 9 and formulated use cases in deliverable D2.2.1 which outline scenarios in which a consumer wants to use or re-use data stored in an archive. The stakeholder and use case definition forms a necessary framework for further analysis of organizational preservation factors.

A number of domain specific collaborations and resource centers have included digital preservation in their program, spreading information about digital preservation best practise in their domain and working on domain specific standard and best practise development. Examples for this are UK Archaeology Data Service (ADS)[115], the Geospatial Data Preservation Resource Center[116] or the Inter-University Consortium for Political and Social Research (ICPSR)[117]. A similar resource center for the architectural domain does not exist - therefore no central point of information for domain relevant community and technology watch resources exists.

The definition of significant characteristics described in chapters 2.6.3 and 3.6.3 as well

[115] http://archaeologydataservice.ac.uk/advice/preservation

[116] http://geopreservation.org/faq.jsp

[117] http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html

as the preservation planning activities based on them require a sound understanding of the object, environment and action alike. On an object level, a definition of content specific technical metadata as mentioned in the previous section 4.4 will assist the process of significant property definition. Stakeholder requirements in the objects should be collected and evaluated through suitable methods like interviews or questionnaires.

Tieing together the information from community watch sources, technology watch sources and stakeholder requirements in a sample preservation plan may function as helpful tool in communicating preservation process needs and factors to the domain.

1. Implementation level of archival practises, i.e., in architectural practise, and preservation knowledge of stakeholders is not clear
2. Stakeholder requirements in curational workflows, i.e., to assess quality of digital objects, is not clear
3. Due to the lack of a central domain specific information resource, no aggregated list of technology watch sources for 3D architectural objects exists
4. Due to the lack of a central domain specific information resource, no aggregated list of community watch sources for 3D architectural objects exists
5. Broad definition of possible significant characteristics at object level is missing due to lack of technical metadata definition
6. No exemplary preservation plan for 3D architectural data is available

5 Conclusion and DURAARK objectives

The goal of this deliverable was to identify gaps in existing processes for the digital preservation of 3D objects. This was achieved through an in-depth analysis of the state of the art for digital preservation in general, regardless of the content-type it targets, and a second in-depth look at projects, existing standards and established processes for the preservation of 3D data. The approach touched on major cornerstones of digital preservation practise: preservation processes for every layer of an object (bit preservation, logical preservation and semantic preservation), organizational processes and metadata as a manifest of a holistic object, content and process description.

While bit preservation can be regarded as solved on a technical level, where risks such as data carrier obsolescence or human error can be countered through refreshing, replication, redundant storage and monitoring, the least developed domain of digital preservation is semantic preservation. Here, projects are just starting out, exploring the possibilities of enriching archives with data available in external sources such as datasets on the web and defining the implications that the long term accessibility of such data has. The biggest advances in digital preservation have been made at the logical level of an object. File format identification, technical metadata extraction and file format validation have been established as good practise within the digital preservation domain. Current projects are either focusing on the scalability of these processes or on the extension towards new content types.

3D data is certainly a content type which has so far not been explored in logical preservation. The study of related projects and guidelines showed that preservation processes at a logical level have only been exploited rudimentarily so far, with a few native CAD formats included in standard preservation format registries such as PRONOM. While there may have formerly been reservations towards IFC-SPF as not having been adopted enough within the architectural domain, the sustainability study conducted in chapter 3.3.2 showed that both IFC-SPF and E57 are well supported, adopted by the community and furthermore open and transparent file formats, making them ideal for digital preservation. Within the DURAARK project the gaps identified in connection with logical preservation will be addressed as part of the ingest process into an existing digital preservation system.

As part of the ingest and storage into an existing digital preservation system, the selection of metadata to be stored alongside the objects plays a key role. While PREMIS is the standard metadata schema for preservation metadata, chapter 3.5 explored some existing recommendations for descriptive metadata. A more in-depth analysis of metadata standards is being conducted in the deliverables D3.3.1 and D3.3.2.

As mentioned before, research in semantic preservation is a relatively new research field in digital preservation.[43] While the KIM project has exploited annotation methods in order to strengthen the semantic information of a digital object over its lifecycle (see chapter 3.1.7), no monitoring of changes within these concepts has been undertaken. DURAARK will semantically enrich objects to be archived with information available on the web, such as e.g., product information of specific building parts. Research questions explored will range from source identification to preserving the datasets used for the enrichment processes.

The stakeholders' needs and requirements will be analyzed through use cases and evaluation of the DURAARK outcomes. They shall lead to a good understanding of the data producers', consumers' and data stewards' requirements in the preservation process. To put it in the words of Stephen Gray of JISC, who said about archiving in general and archiving of 3D in particular: "Archiving is far easier when repositories (museums, archives, universities or similar) work alongside the creators of 3D content with preservation as the common goal" [20].

References

[1] I. Ahmed, K.-s. Lhee, H. Shin, and M. Hong. On improving the accuracy and performance of content-based file type identification. In C. Boyd and J. González Nieto, editors, *Information Security and Privacy*, volume 5594 of *Lecture Notes in Computer Science*, pages 44–59. Springer Berlin Heidelberg, 2009.

[2] ANSI/NISO. ANSI/NISO Z39.87 - Data Dictionary - Technical Metadata for Digital Still Images.

[3] D. Arnold. 3D-COFORM: D.3.3 - Third Year Report. WP3 - Repository Infrastructure. November 2011.

[4] D. Arnold. D.1.5. - Project Final report. 3D-COFORM Project. 20 September 2013. 2013.

[5] T. Austin and J. Richards. Archaeology Data Service: Preservation Policy, September 2009.

[6] A. Ball. Review of data management lifecycle models. 2012.

[7] A. Ball. Preserving Computer-Aided Design (CAD). Technical Report 13-02, DPC, April 2013.

[8] C. Becker, S. Strodl, R. Neumayer, A. Rauber, E. Nicchiaerlli Betteli, and M. Kaiser. Long-term preservation of electronic theses and dissertations: A case study in preservation planning. In *Proceedings of the Ninth Russion National Research Conference on Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL'07)*, 2007.

[9] I. Blümel. Documentation of PROBADO3D metadata for OAI-PMH interface. October 2012.

[10] S. Boeykens and E. Bogani. Metadata for 3D Models. How to search in 3D Model repositories? In *ICERI 2008 Proeceedings*, 2008.

[11] A. Brown. Selecting File Format for Long-Term Preservation. Version 2. Digital Preservation Guidance Note 1., 2008.

[12] P. Caplan. Understanding PREMIS. Technical report, Library of Congress Network Development and MARC Standards Office, February 2009.

[13] CCSDS. Reference Model for an Open Archival Information System (OAIS) - Magenta Book, 2012.

[14] A. Dappert and A. Farquhar. Significance Is in the Eye of the Stakeholder. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Research and Advanced Technology for Digital Libraries*, volume 5714 of *Lecture Notes in Computer Science*, pages 297–308. Springer Berlin Heidelberg, 2009.

[15] M. d'Aquin, A. Adamou, and S. Dietze. Assessing the educational linked data landscape. In *WebSci*, pages 43–46. ACM, 2013.

[16] H. E. Denard. The London Charter for the Computer-Based Visualisations of Cultural Heritage. February 2009.

[17] S. Dietze, J. Beetz, U. Gadiraju, G. Katsimpras, R. Wessel, and R. Berndt. Towards Preservation of Semantically Enriched Architectural Knowledge. In *SDA*, pages 4–15, 2013.

[18] M. Doerr and M. Theodoridou. CRMdig: A generic digital provenance model for scientific observation. In *Proceedings of TaPP'11: 3rd USENIX Workshop on the Theory and Practice of Provenance*, 2011.

[19] B. Fetahu, S. Dietze, B. Pereira Nunes, D. Taibi, and M. A. Casanova. Generating structured profiles of linked data graphs. In *The Semantic Web - ISWC*, 2013.

[20] S. Gray. Building 3D content to last. JISC Digital Media, 2012.

[21] H. U. Heidbrink. Long term preservation of engineering data in the shaman project. the industrial use case within the shaman project (isp-2). Presentation held at the DPC Preserving CAD Briefing Day., July 2013.

[22] S. Higgins. The DCC Curation Lifecycle Model. *The International Journal of Digital Curatoin*, 3(1):134–140, 2008.

[23] D. M. Jones, editor. *3D Laser Scanning for Heritage.* English Heritage, second edition edition, October 2011.

[24] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogen. Oberving linked data dynamics. In *The Semantic Web: Semantics and Big Data*, pages 213–227. Srpinger Berlin Heidelberg, 2013.

[25] T. Kuny. A digital dark ages? challenges in the preservation of electronic information. In *Proceedings of the 63rd IFLA Council and General Conference*, 1997.

[26] Library of Congress. Sustainability of Digital Formts - GIF Graphics Interchange Format, Version 89a. Webpage, 2013.

[27] J. Ludwig. About the complexity of a digital preservation theory and different types of complex digital objects. Dagstuhl Seminar 10291, 2010.

[28] J. Mitcham. *The Preservation of Complex Objects, Volume 1: Visualisations and Simulations*, volume Volume 1: Visualisations and Simulations, chapter Preservation of Digital Objects at the Archaeology Data Service, pages 76–85. The University of Portsmouth, 2012.

[29] M. Multimedia. *Standards Update: New Multimedia Data Types and Data Techniques.* Microsoft Corportion, 1994.

[30] NISO. Understanding metadata, 2004.

[31] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM, 2004.

[32] M. Patel and A. Ball. Strategies for the Curation of CAD Engineering Models. *The International Journal of Digital Curation*, 4:84–97, 2009.

[33] M. Patel, A. Ball, and L. Ding. Curation and Preservation of CAD Engineering Models in Product Lifecycle Management. In *Digital Heritage - Proceedings of the 14th International Conference on Virtual Systems and Multimedia*, 2008.

[34] E. Peters McLellan. General study 11 final report: Selection digital file format for long-term preservation. Online, March 2007.

[35] PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata, version 2.2, 2012.

[36] S. Project. *D14.3 - Report on demonstration and evaluation activity in the domain of industrial design and engineering. SHAMAN - WP14-D14.3.* 2011.

[37] J. Rog and C. van Wijk. Evaluating file fformat for long-term preservation. National Library of the Netherlands; The Hague, The Netherlands, 2008.

[38] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272(1):42–47, 1995.

[39] P. Rula, M. Palmoari, A. Harth, S. Stadtmüller, and A. Maruino. On the diversity and availability of temporal information in linked open data. In *The Semantic Web - ISWC 2012*, pages 492–507. Springer Berlin Heidelberg, 2012.

[40] C. Schlieder. Digital heritage: Semantic challenges of long-term preservation. *Semant. web*, 1(1,2):143–147, Apr. 2010.

[41] B. Sierman, C. Jones, S. Bechhofer, and G. Elstrøm. Preservation Policy Levels in SCAPE. In *Proceedings of the 10th International Conference on Preservation of Digital Objects iPRES 2013*, 2013.

[42] M. Smith. Final report for the mit facade project: October 2006 - august 2009. Massachusets Instituite of Technology, 2009.

[43] S. Strodl, P. Petrov, and A. Rauber. Research on digltal preservation within project co-funded by the european union in the ict programme, May 2011.

[44] D. Taibi, B. Fetahu, and S. Dietze. Towards integration of web data into a coherent educational data graph. In *WWW (Companion Volume)*, pages 419–424, 2013.

[45] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web abd web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 6:165–182, 2011.

[46] K. Thibodeau. *The State of Digital Preservation: An International Perspective*, chapter Overview of Technological Approaches to Digital Preservaton and Challanges in Coming Years, pages 4–31. CLIR, 2002.

[47] UK Data Archive. *Preservation Policy*, 2012.

[48] J. van der Knijff and C. Wilson. Evaluation of characterisation tools. part 1: Identification. September 2011.

[49] E. Zierau. *A Holistic Approach to Bit Preservation.* 2011.