The DURAARK Project – Long-Term Preservation of Architectural 3D-Data

Authors: Michelle Lindlar, M.A. / Dr. Hedda Saemann (TIB/UB Hannover)

Contact:

Technische Informationsbibliothek und Universitätsbibliothek Hannover (TIB/UB)
Welfengarten 1B
30167 Hannover
Germany

Michelle.Lindlar@tib.uni-hannover.de
Hedda.Saemann@tib.uni-hannover.de

Introduction

For centuries two-dimensional presentations of objects have been predominant in the depiction of architecture. In building research traditional methods still establish the basis supported by geodesic instruments – like the tachymeter – as an integral part of fieldwork, whose results will then be transferred into CAD plans by means of point clouds. As digital technologies are gradually taking over, these two-dimensional point clouds and drafts are completed by three-dimensional methods of recording and presenting. In this way the preservation of scientific and cultural heritage is increasingly linked to the maintenance of the long-term availability of digital information, for which no general strategy has been developed yet. In the cultural heritage domain, a digital representation may exist as a documentary representation of an existing object. However, in some cases, it may also be the only remaining documentation of an (analogue) object which has been destroyed or lost.

Thus, on the one hand, a mixture of analogue materials, like notes and sketches, as well as "born digitals", like CAD drawings and laser-scans, have to be coped with today. On the other hand, technological problems have to be solved, e.g. the aging of hardware or proprietary data formats. Another, not less alarming problem is constituted by cultural oblivion, e.g. information may have been omitted when documenting the object, because it was taken for granted, or considered part of the knowledge of the contemporary target-groups.

However, to what extent do such problems concern institutions like the TIB? The German National Library of Science and Technology – TIB – functions as a national library for the subject matters of engineering, architecture, chemistry, computer science, mathematics and physics.[1] It hosts the cooperatively operated digital preservation system of *Goportis*,[2] a consortium consisting of the German National Subject Libraries.

TIB is responsible for the procurement and archival storage of specialized technical scientific literature from all over the world, providing the national and international research community and the industry with in-depth information. In addition to textual documents TIB is engaged in many different kinds of media such as research data, software, as well as scientific films, simulations and 3D-objects.

Until recently the focus of digital preservation has mainly been on text and still images and state of the art knowledge has yet to be applied to other materials, like architectural data. Filling gaps in architectural data is the assignment of the DURAARK project with the participation of the Leibniz University of Hanover, closely affiliated with TIB.

DURAARK – DURAble ARchitectural Knowledge

DURAARK is an EU-funded research project which started in February 2013 and will be concluded in January 2016. The multilayered research assignments require expert knowledge from diverse

[1] http://www.tib.uni-hannover.de/en/tibub/about-us/library-profile.html
[2] http://www.goportis.de/en/about-goportis.html

fields. In the interdisciplinary DURAARK project consortium this is gathered from the fields of architecture and civil engineering (partners: CITA - Center for Information Technology and Architecture/Denmark; Catenda/Norway; Department of the Built Environment at the Technical University of Eindhoven/Netherlands), visual computing (partners: Department of Computer Graphics at the Universität of Bonn/Germany; Department of Visual Computing at Fraunhofer Austria Research GmbH FhA), semantic web (Research Centre L3S, Hannover/Germany) and long-term digitization (partners: Luleå University of Technology, Sweden; Technische Informationsbibliothek Hannover/Germany).

The three-year project aims at the development of methods for the sustainable long-term preservation of building data, which include 3D point cloud scans and building information models (BIM) as well as metadata, related knowledge and Web data. From the ingest, storage and retrieval of 3D objects to measurements for the preservation of long-term availability, DURAARK covers a wide range of processes and methods especially tailored to the domain of architectural 3D content. Its scope covers a holistic digital preservation approach as well as interlinked curation and preservation workflows for data ranging from low-level 3D point clouds to highly annotated 3D BIM models. Two different ways of enrichment are being explored in order to ensure the long-term interpretability of content: semantic and geometric enrichment. The storage of the semantically and geometrically enriched objects is to be achieved by incorporating an existing OAIS compliant digital and preservation system at TIB.

Architectural 3D object types – scans and plans

As mentioned, DURAARK deals with two different types of architectural data: first 3D scans (point clouds) which can be captured with various scanner types, such as hand-held devices or drones. The objects are a set of points in a 3D coordinate system which describe the external surfaces of a scanned object. As they document a building or structure at a fixed point in time ("as-is"), they are "descriptive representations" inevitably tied to temporal and spatial aspects.

The second object type, the (BIM), goes significantly further. Based on different 3D modelling processes in use since the 1980s, the capacities of 3D plans employed in architecture were steadily developed – up to the multispectral depiction of diverse processes of design and execution. Thus, 3D models no longer only serve as visual aids, as analogue, true-to-scale models have for centuries. BIM objects move beyond CAD-geometry by covering the entire design-to-construction process, including aspects such as schedule time (4D), cost-related information (5D), energy and sustainability (6D) and facility management (7D). They traditionally document a building or structure "as-planned", making them "prescriptive representations", which may deviate from the "as-is" state.

To limit the scope of the project, one file format was chosen for each of the two object types: the ASTM[3] standardized E57 file format for point-cloud scans and the ISO standardized[4] IFC (Industry Foundation Classes) for BIMs. In line with digital preservation best practices, the sustainability of those two file formats was carefully considered. Both are open standards, which are well documented, embraced by community and industry and transparent to digital preservation processes.

Long-term usage scenarios for architectural data

In a building project, the inclusion of the different crafts and trades involved and of the data recorded for and by each of them (vendor information) creates an enormous mass and variety of data about one object, which is used to enhance the object as a whole, but also each of its single elements. As a result, BIM is able to generate diverse views and plans for the different crafts, e.g. the facade, the structure or the technical equipment and appliances. Thus the shift from analogue 2D

[3] http://www.astm.org/COMMITTEE/E57.htm
[4] http://www.iso.org/iso/catalogue_detail.htm?csnumber=51622

plans to digital 3D representations creates a growing data pool ideally documenting an object's whole life cycle. Such complex time studies can be irreplaceable cultural heritage objects themselves.

A major challenge to the long-term preservation of architectural data is constituted by the long periods of time during which the data may be dormant. This is a direct result of the usual long operating period of a building, during which data may be accessed seldomly, until a major retrofit event occurs every 20-30 (or more) years. As regards technological change, this represents an incalculable span of time. And yet, from the perspectives of cultural heritage and historical building research, which rather count by centuries, access every three decades may seem "frequent".

These developments affect libraries and archives for several reasons: First, 3D scans and plans are the research output of their respective departments and will be found in future legacy deposits of architects and engineers as well as in current institutional repositories. Second, building information modeling (BIM) is already mandatory in a number of countries for publically funded buildings, including Denmark, Finland, Hong Kong, the Netherlands and Norway, Singapore, the UK and the USA. In addition, of those all but Hong Kong require the IFC format for BIM, which is one of the factors that have led to the choice and enhancement of this format in the DURAARK project [1]. Last but not least, cultural heritage organizations are, of course, already scanning (and also planning) cultural heritage sites for documentation.

As the cultural heritage interest in documentary digital objects of a structure may far surpass the actual lifecycle of the erected structure itself, cultural heritage institutions should be regarded as a separate consumer, to whom the preservation of the current state of a building is as important as that of records of older stages – for historical value. Although interests may overlap here, researchers form a separate stakeholder group, which solemnly functions as a designated community with different knowledge and different demands as to the data's appropriateness and completeness. Researchers querying the data in connection with e.g. socio-economical or historical research questions may add new sets of knowledge to the digital object [2].

A holistic digital preservation view

As digital preservation practises have matured, it has become clearer that "one-size-fit-all solutions" do not exist. Many libraries and archives have implemented "one-size-fit-many solutions", such as workflows normalizing textual data to PDF/A, however, these workflows will certainly not fit the preservation intent and the preservation need of all future users and are not suited for all types of content. While digital preservation has been largely driven by cultural heritage institutions for the past 15 years, Strodl et al. have noticed a slow awareness shift in various domains from where the data originates [3]. This is especially true for non-textual materials as well as for "living" process data. 3D architectural data fits into both of those categories.

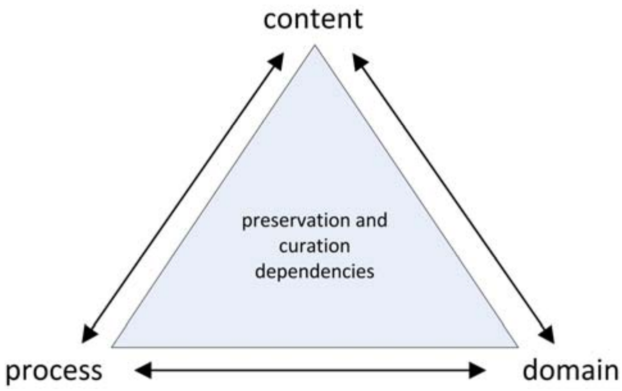

Figure 1: (Digital) preservation and curation dependencies

Preservation and curation strategies depend on three areas: the **content** to be preserved, which stems from and is preserved for a **domain**[5], who in return uses a domain-specific **process** to create, access and re-use the content (see figure 1).

While the most common framework to design, control and discuss digital preservation processes within an organisation is the OAIS (Open Archival Information System) model, further models are frequently used to explain these processes on a more granular level. The DURAARK project has combined several models into a holistic digital preservation and curation view shown in figure 2. While the outer layer is based on expectations of the domain(s) in their roles as "producer", "archive" and "consumer", the inner circle focuses on the processes in which the objects are created, modified, accessed and re-used. This layer of the model is based on the DCC (Digital Curation Center) Curation Lifecycle Model [4]. At the core of the model is the digital object itself, which is not regarded as a monolithic entity, but seen strictly in the context of the three digital object layers as formulated by Thibodeau: the bit-stream layer, the logical layer and the semantic layer [5].

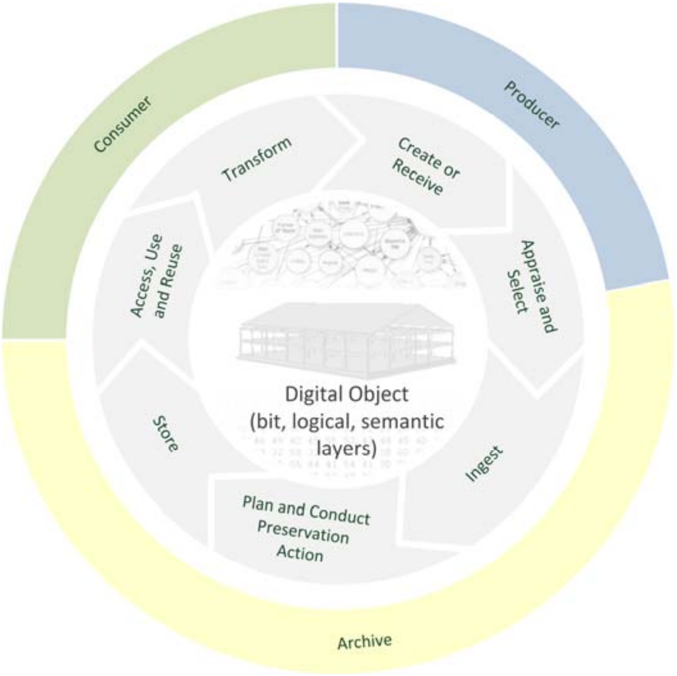

Figure 2: Holistic digital preservation and curation view

Identified gaps and corresponding DURAARK tasks

One of the first tasks of the DURAARK project was an analysis of the state of the art of digital preservation in general as well as digital preservation of 3D objects in particular. The comparison of the two areas allowed the project consortium to derive concrete gaps, which are currently being addressed in the remainder of the project time. This analysis followed the different layers of the holistic preservation model (see figure 2), starting with the three layers of the digital object (bitstream, logical, semantic) and moving from there on to process and stakeholder requirements and limitations [6].

Typical risks associated with the bitstream layer are media obsolescence, technical failure, digital rights management measures which hinder preservation processes or human error, such as accidental deletion. Typical actions taken to counteract these risks are media migration, refreshing and replication, error detection and monitoring processes as well as a system based on technological

[5] It is important to note that the producing domain and the consuming domain do not have to be the same, which leads to different expectations in the material.

redundancy, ideally with a geographic spread. None of these risks are unique to architectural 3D objects and all are solved through "good IT practise" – which, of course, needs to be implemented. In support of this, trustworthy certification processes, like the Data Seal of Approval[6], specifically require organizations to document their data storage processes.

Typical risks associated with the logical layer are connected to the file format and its dependency on software and with that on an operating system and potentially on a hardware platform. This also includes dependencies on certain packages – as in the case of non-embedded fonts – or on configuration, as in the display of measurements without accompanying information as to their unit. As described earlier, the file formats E57 and IFC were analyzed in the context of their sustainability factors – an important task in regard to logical preservation.

Standard digital preservation processes pertaining to the logical layer are the characterization of the object in a first step, which includes the identification of the file format, the validation of the object against a file format specification and the extraction of technical metadata. Technical metadata is often not specific to a file format, but to a content type – as in the case of the NISO Metadata for Images in XML (MIX) standard, the de-facto technical metadata standard for still images.[7] It describes representation criteria beyond the file format layer, such as the encoding of the data stream and spatial and/or temporal aspects.

The characterization of digital objects upon ingest into an OAIS compatible archive and the capturing of the results in managed metadata form the pre-requisites for any preservation action. While the two open formats chosen for the DURAARK object types – E57 for point-cloud scans and IFC-SPF for BIM objects – are well suited for archival purposes from a sustainability point of view, a number of gaps regarding the characterization process were identified. The project closed the gaps regarding lacking file format identification with both E57 and IFC now included in the TNA (The National Archives, UK) operated file format registry *PRONOM*[8] and identifiable through the widely used file format identification tool *DROID*[9]. The project is currently evaluating whether the identification of IFC needs to be further broken down into its schema variants. With file format identification in place, the definition of technical metadata set for architectural 3D data is a step currently being tackled by the project. As point clouds and BIM both represent architectural 3D objects but substantially differ from another on a representational level, different sets of technical metadata need to be defined for the content types.

First candidates identified for *e57m* technical metadata included RGB colour values and the number of scans; first candidates identified for the *ifcm* technical metadata subset included the percentage of optional schema-level attributes and the last software to have modified the file [7]. A definition of a full first draft of technical metadata attributes for the two sets is currently under development, alongside the development of tools and methods to extract these values from the digital objects.

File format validation is the last important building block of file format characterization. For both formats, validation tools are already available. While the *libE57* reference implementation[10] tool *e57validate* produces stable results against the file format specification, IFC validation is a much harder process. As IFC-SPF files are encoded at structured ASCII files the validation of the basic level is straightforward – however, when taking validation down to the schema instance, IFC is too complex. The latest schema IFC4 consists of 766 entities with approximately 2000 schema-level attributes. In addition to these, hundreds of standardized properties can be assigned to each entity and can furthermore be extended to suit the needs of end-users and software vendors alike [8]. In the context of the project the DURAARK consortium is in the progress of evaluating the feasibility of schema subsets for an archival extension of IFC.

The last layer of the digital object is the semantic layer. Here, typical risks are associated with terminology and concepts which change over time as well as with context and provenance which

[6] http://www.datasealofapproval.org/

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

[8] http://apps.nationalarchives.gov.uk/PRONOM

[9] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

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

may be lost over time. Concrete examples from the architectural world are those of changing street or city names or repurposing of buildings while their names may still refer to their old functions.

The DURAARK project is tackling the preservation problems on this object level through semantic enrichment of the object. For this, a “semantic digital archive” (SDA) is under development, which allows a curator to semantically enrich an architectural 3D object. The SDA serves as a central knowledge base for contextual knowledge in the form of structured data about architectural structures. Examples of datasets contained in the SDA include vendor databases, *Geonames* or *DBPedia*. Further datasets to be included in the SDA are being identified.

The SDA furthermore contains a snapshot of the enriched data. This is of special relevance in the case of vendor information, often merely referenced from within an architectural digital object and therefore not persistent information archived with the object itself. During the enrichment process the archivist can replace external links to links existing within the SDA, guaranteeing the long-term availability of the semantic information itself. The SDA is accompanied by a semantic digital observatory (SDO), which profiles external datasets and is responsible for monitoring and capturing new snapshots of existing datasets or new datasets altogether [9].

An important part in maintaining contextual and provenance information over a long time is, of course, structured metadata. As several existing metadata schemas either focus on the description of the physical structure (e.g. *MIT Façade* PIM model), describe only a specific digital content type (e.g. Historic Buildings and Monuments Commission for England) or inversely lack content-specific information (e.g. *CARARE* or *PROBADO 3D* metadata core), DURAARK is proposing the *buildm* schema, which describes both the physical object as well as the data object. It also differentiates between different content types within the data object section (see figure 3). The metadata requirements are closely tied to stakeholder requirements and limitations.

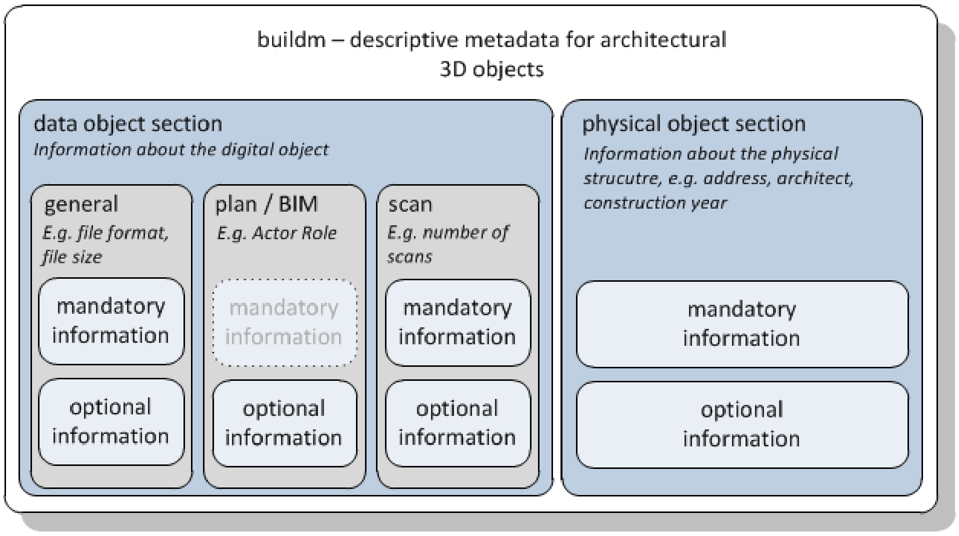

Figure 3: Draft of the buildm metadata schema

The role of stakeholders

The preservation dependencies (figure 1) shows a direct connection between content, processes and the domain. As one of the project’s first tasks, different stakeholders with a potential interest in long-term availability and interpretability of architectural 3D data were identified and mapped so that the “domain” could be replaced by three actor roles, namely the “producer”, the “archivist” and the “consumer” (see figure 2: the holistic preservation view). Any processes and methods developed within the scope of the DURAARK project therefore have to be matched to these three groups of stakeholders [2][6][10].

In a second step, the processes as part of which the data is created, used and re-used were analyzed with regard to the different stakeholders. One of the key findings of this in-depth analysis regards the long-term archiving incentive amongst the stakeholders. While especially building owners, public administrators, cultural heritage institutions and researchers showed a strong interest in long-term availability and usability of the data, the same did does not hold true for the AEC (architecture, engineering, construction) domain itself. The majority of architects, 3D scanning companies and engineers interviewed showed no to little interest in the long-term availability of their own data output. Exceptions to this general observation were a short-term interest for archiving amongst construction companies who need to keep the data available for construction-site reviews, e.g. in the case of Scandinavian 5 year building reviews [11].

DURAARK use cases

In order to understand the incentives for the different stakeholders to archive data, a number of concrete use cases to be addressed within the project were gathered [10]. The use cases can be separated into 4 classes:

Class 1: Meta-use cases for the integration of DURAARK results into an OAIS compliant archive

- *Deposit 3D architectural objects*
- *Search and retrieve archived objects*

Class 2: Meta-use cases for the development of an implementation of a Semantic Digital Archive

- *Maintain the semantic digital archive*
- *Enrich a BIM/IFC model with metadata form a repository*

Class 3: Curational use cases on the geometric level

- *Detect differences between planning and as-built state*
- *Monitor the evolution of a structure over time*
- *Identify similar objects within a point-cloud scan*

Class 4: Curational use cases on the semantic level

- *Plan, document and verify retrofitting/energy renovation of buildings*
- *Exploit contextual information for urban planning*

While classes 1 and 2 are being addressed by the DURAARK tasks which were mentioned earlier on in this paper and are located at the ingest stage or within the archive itself, classes 3 and 4 describe use cases in which a consumer wants to access the data and re-use it for a specific purpose. These use cases, in return, rely on available tools for execution but also indirectly formulate requirements for the geometric or semantic information available within or alongside the digital object to be preserved.

The DURAARK project is developing tools to support all of the cases of use – in the case of curational use, the implications this has for the object, metadata and process requirements are noted and realized in the processes that support the preservation process itself.

Conclusion and outlook

In the first half of its running time, the DURAARK project has identified gaps in the digital curation and preservation processes of architectural 3D data and begun to close these. The basis for the identification is a holistic digital preservation view, which considers different digital object layers as well as procedural and actor-based requirements, capabilities and limitations. First tools and methods to support curational use such as the comparison of a point-cloud scan and a BIM (in order to detect differences between planning and "as-is" state) are now in place [12], and an ontological framework for the semantic digital archive has been drafted [9]. The project is currently in the process of developing a first integrated prototype, which will support the use case classes 2-4. The main outcome of the prototype are “archival-ready SIPs” (submission information packages) – a necessary pre-requisite for the deposit into an OAIS-compliant archive.

The next step of the project is the evaluation of the intermediate results through the stakeholders. Here, interviews and workshops will be conducted to measure the alignment of the current results with the stakeholder expectations. Based on the outcome, the processes will be adjusted accordingly. The project results are being made available via the project’s website: http://www.duraark.eu.

Acknowledgements

This work was partially funded by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600908 (DURAARK) 2013-2016.

References

[1] Michelle Lindlar. Building Information Modeling – A Game Changer for Interoperability and a Chance for Digital Preservation of Architectural Data? Forthcoming in: iPRES 2014. Proceedings of the 11th International Conference on Digital Preservation. (2014)

[2] Michelle Lindlar, Martin Tamke. A Domain-driven Approach to Digital Curation and Preservation of 3D Architectural Data: Stakeholder Identification and Alignment in the DURAARK Project. In: Society for Imaging Science and Technlology.Archiving 2014 Final Program and Proceedings. pp. 204-209. (2014)

[3] Stephan Strodl, Petar Petrov, Andreas Rauber. Research on Digital Preservation within projects co-funded by the European Union in the ICT programme. (2011)

[4] Sarah Higgins: The DCC Curation Lifecycle Model. The International Journal of Digital Curation, Issue 3, pg.134-140. (2008)

[5] Kenneth Thibodeau. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In: The State of Digital Preservation: An International Perspective. CLIR Report 107. (2002)

[6] DURAARK. Current state of 3D object digital preservation and gap-analysis report. Deliverable Report D6.6.1 (2014)

[7] DURAARK. Meta data schema extension for archival systems. Deliverable Report D3.3.1. (2014)

[8] BuildingSMART. Industry Foundation Classes Release 4 (IFC4). Specification Documentation. *(*2013).

[9] DURAARK. Ontological Framework for a Semantic Digital Archive. Deliverable Report D3.3.2. (2014)

[10] DURAARK. Requirements Document. Deliverable Report D2.2.1. (2013)

[11] DURAARK. Current state of 3D object processing in architectural
research and practiceDeliverable Report D7.7.1. (2014)

[12] DURAARK. Documenting the Changing State of Built Architecture – Software prototype v1. Deliverable Report D4.4.1