www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

UNVEILING BIM CORE: ENHANCING LONG-TERM USABILITY OF DIGITAL BUILDING DOCUMENTATION WITH AN ADVANCED REPRESENTATION INFORMATION REPOSITORY

SUBMITTED: November 2023
REVISED: July 2024
PUBLISHED: August 2024
EDITOR: Bimal Kumar
DOI: 10.36680/j.itcon.2024.027

Uwe M. Borghoff, Professor,
Dept. of Computer Science, Inst. for Software Technology, University of the Bundeswehr Munich, Germany
ORCID: https://orcid.org/0000-0002-7688-2367
uwe.borghoff@unibw.de (*corresponding author)

Eberhard Pfeiffer, Academic director,
Dept. of Civil Engineering and Environmental Sciences, University of the Bundeswehr Munich, Germany
ORCID: https://orcid.org/0000-0001-5074-5797
eberhard.pfeiffer@unibw.de

Peter Rödig, Researcher/Curator of the datArena,
Dept. of Computer Science, Inst. for Software Technology, University of the Bundeswehr Munich, Germany
ORCID: https://orcid.org/0000-0002-1947-5909
peter.roedig@unibw.de

SUMMARY*: The long-term usability of digital building documentation is essential for the maintenance and optimization of infrastructure portfolios. It supports the preservation of building-specific knowledge and the cultural heritage hidden within. However, having to do this throughout the entire lifecycle of a building—or even indefinitely—remains a major challenge. This is especially true for organizations responsible for large collections of digital building documents and datasets, such as public administrations or archives. In this article, we first describe the challenges and requirements associated with preservation tasks and then introduce the concept of representation information within Building Information Modeling (BIM) and all types of related data and documents. This type of information is important to give meaning to the stored bit sequences for a particular community. We then design a repository for representation information and propose some 23 so-called BIM Core content elements. Finally, we focus on BIM and the construction sector and explain how the proposed repository can be used to implement the two concepts introduced in the ISO reference model Open Archival Information System (OAIS), namely the representation information and the context information, as well as the concept of significant properties, which has not yet been explicitly modeled in OAIS.*

KEYWORDS*: long-term archiving, representation information repository, digital building documentation, building information modeling (BIM), concept of significant properties, BIM Core content elements.*

REFERENCE*: Uwe M. Borghoff, Eberhard Pfeiffer, Peter Rödig (2024). Unveiling BIM Core: Enhancing Long-Term Usability of Digital Building Documentation with an Advanced Representation Information Repository. Journal of Information Technology in Construction (ITcon), Vol. 29, pg. 596-611, DOI: 10.36680/j.itcon.2024.027*

1. INTRODUCTION

Resource scarcity, cost increases, climate change, and demographic change pose major challenges for the public and private construction sector. However, digitization of these sectors can help to manage the associated tasks, especially if the entire life cycle of buildings, including demolition and recycling of materials, is considered on the basis of digital information, or digital twins in the future (Alshammari et al., 2021, Madubuike et al., 2022).

Our project partner, the Bavarian Building Authority, analyzed the value of its digital assets and the cost of subsequent digital recording in 2015. The result was an average of about 3 euros per square meter of gross floor area (GFA), based on 16,320 buildings with a total of 27,724,000 square meters of GFA. It is therefore well worth taking a closer look at the benefits and challenges.

1.1 Benefits and Challenges

There is no doubt that information and communication technologies (ICT) have had a positive impact on the construction sector in general, and the use of ICT has led to significant improvements in productivity and efficiency (Chen et al., 2022). Its integration into the construction industry has also been a driver of sustainability.

Digital tools are helping to improve the management of resources and energy efficiency. For example, waste and the environmental footprint of construction activities can be reduced by using precise measurements and data analytics to ensure that materials are used optimally (Weili et al., 2022). In particular, the ability to use digital building information on an ongoing basis can provide the following concrete benefits:

- The avoidance of the loss or degradation of information, and the repeated and cost-intensive creation of new data in the case of renovations, repurposing, repairs, deconstruction planning or waste management (Ge et al., 2017).
- The preservation and expansion of knowledge for new construction projects to avoid reinventing the wheel or making the same mistakes over and over again; the more information about the context of a building that can be taken into account, the better this will work. Zahedi et al. (2022) presented an approach to link client requirements and building codes to design concepts, and to record and document design decisions and their explanation in a way that is transparent to all stakeholders to improve knowledge capture.
- The preservation and expansion of knowledge for use in new (ICT) projects, especially in data and information management technology, communication and visualization, and automation and analytics (Toyin et al., 2024).
- The economic implementation of complete long-term analyses of buildings and the economic viability of building portfolios (Forum, 2024) as well as appropriate organizational and business models (Beinert et al., 2008).
- The improvement of computational models or their parameters on the basis of data from long-term monitoring (Panah and Kioumarsi, 2021).
- The planning and simulation of safety and emergency response through real-time monitoring and risk management at every stage of the lifecycle, from design and construction to operation and decommissioning (Waqar and Ahmed, 2023).
- The preservation of cultural built heritage through virtual structures, especially when physical preservation or restoration is not possible (Li et al., 2023).
- The fulfillment of all formal requirements for documentation and archiving, especially the development of evaluation grids (Leventhal et al., 2021).

Ensuring the long-term usability of digital information and safeguarding the digital cultural heritage is one of the grand challenges of information technology (GI, 2014). A UNESCO publication highlights the need for digital preservation to protect cultural heritage, emphasizing that maintaining the long-term accessibility and usability of digital resources is a critical issue that requires careful consideration and strategic planning (Choy et al., 2016). The particularities and constraints of the construction and public administration sectors add to the challenges. Cursi et al. (2022) analyze the extension of the BIM process to extend the semantic scope and modeling capabilities of today's market tools to meet the specific needs of built heritage applications.

A number of long-standing issues can be identified that relate specifically to the objectives of this paper, namely

- the rapid obsolescence of formats due to short innovation cycles for hardware, software, and human-machine interaction (Borghoff et al., 2006);
- the large number of formats for data, files, file systems, database schemas, protocols for data exchange as well as the insufficient documentation of their context (Halevy, Rajaraman, and Ordille, 2006) although some improvements have been made (Golshan et al., 2017);
- increasing division of labor and distribution of related storage and processing systems, especially with respect to current BIM data exchange methods (Lou et al., 2021);
- incomplete standardization, proprietary formats, and imprecise specifications, especially regarding trustworthiness (Dobratz et al., 2010).

The last point in particular causes a lot of problems. Digital collections, especially in public administrations, still contain digital objects in proprietary, outdated or semantically deficient formats, such as plot files or TIFF files, which represent retro-digitized analog plans. This can lead to a number of other challenges and issues, including the fact that TIFF files, while widely supported, can sometimes be saved with proprietary settings or extensions that are not easily accessible or usable without special software. Although the TIFF files themselves are not necessarily obsolete, the software or settings used to create them may become obsolete. As a result, the files may not be accessible or rendered correctly in the future as technology evolves (Hodge and Anderson, 2007).

Government agencies may also have various regulatory requirements for document retention, accessibility, and authenticity (Barton, 2006). Retro-digitized plans must be in compliance with these requirements, which can be another challenge if the digital formats are not well managed or documented.

1.2 BIM and the Digital Transformation

Before we get into today's BIM and what is now known as digital transformation, let's take a historical look at how it all started. The long history of electronic data processing began in the early 1940s with the solution of construction and mechanical problems, and it has been a long road to today's state of the art in building computing. After around 1950, when the first freely programmable computing systems developed by the civil engineer Konrad Zuse became available, it was initially a question of finding out how traditional manual calculation methods tailored to slide rules could be replaced by computer-adapted procedures. The increasing availability of computer capacity at universities and large-scale research facilities led to the rapid development of new numerical methods worldwide, with which complex structures and systems of equations with several tens of thousands of unknowns could be calculated reliably and quickly despite limited computer resources.

Human interaction with computers was long time limited by punched tapes, punched cards, strictly formatted and therefore error prone input formats for commands and data, and often long printouts that had to be laboriously evaluated. Terminals, later with graphical functions and thumb wheels or joystick for cursor operations, digitizer tablets and plotters gradually made work more efficient. Figure 1 shows such a traditional workplace in an IT-laboratory, as it probably existed in a similar form worldwide at the end of the 1970s, consisting of a tablet for digitizing technical drawings and a menu field for the specification of datasets and for program control with symbolic buttons. The configuration shown also included a vector-based display with hardcopy unit, a pen plotter with a tape unit for recording graphic output or possibly input of saved drawings, and still a punch card-reader and a high-speed printer. The computer was a SIEMENS Process 330 system with an alphanumeric console and units for magnetic tapes and removable disks. A FORTRAN compiler and the Tektronix PLOT-10 graphical software served as the development platform for an advanced software library for structural and fluid dynamic analysis. All program developments from this time and to a large extent also from earlier software projects were continuously documented with flow charts, memory plans and source codes. This made it much easier and faster to later rewrite it in other programming languages and to port it to almost all computer platforms, from mainframes to embedded or mobile systems. This is a best practice example of how changing development teams over time could benefit from well-documented prior work when reusing code.

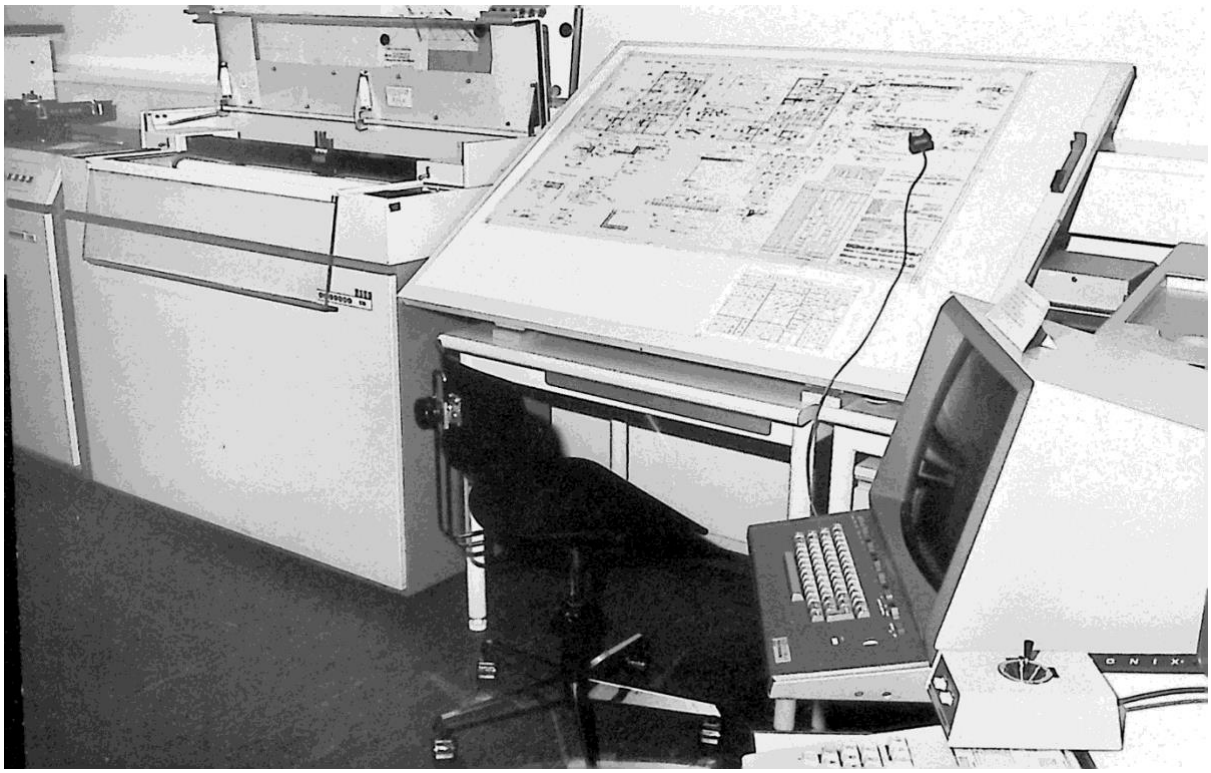

Figure 1: Civil engineer workplace around 1978.

Back then, construction documents were usually paper-based, with computer-generated drawings and extensive lists attached, but in some cases digital media such as floppy disks with input and output files were also included. Because the systems were expensive, not specific enough, and people preferred to work the traditional way, the adoption of CAD systems in the construction industry has been a slow and sluggish process. This changed with the advent of inexpensive minicomputers and graphics-capable devices that made CAD affordable to smaller engineering and architectural firms.

The introduction of IBM personal computers and the first version of Autodesk's popular CAD system, AutoCAD, in the early 1980s, followed by graphical user interfaces a few years later, was a huge boost and driving force for the digital transformation. Applications for AEC (Architecture, Engineering, Construction), MEP (Mechanical, Electrical, Plumbing), HVAC (Heating, Ventilation, Air Conditioning), and project bidding, awarding, and invoicing began to emerge for the diverse trades typical of the construction industry. However, this development was accompanied by a significant increase in the number of different, mostly proprietary formats. Exchanging data and ensuring its consistency and integrity became increasingly difficult.

With the introduction of the first version of the so-called Industry Foundation Classes (IFC) in 1996 and a period of consolidation in the following decade, but at the latest with the ISO standardization (ISO, 2024), BIM has emerged as the single and preferred method for planning and managing construction projects based on a virtual representation of a building with all the features required for integrating the specific models of all the disciplines involved. However, the introduction of BIM brings additional challenges for long-term data and document management.

Advanced analysis methods based on digital building models and innovative monitoring and production models (e.g. virtual spare parts based on 3D printing formats) lead to highly specialized formats and in some cases to large amounts of data. In the context of preservation (e.g. cultural heritage), extended data structures for buildings and their semantic enrichment, such as linked data or discipline-specific ontologies, need to be considered (Pocobelli et al., 2018). In addition, the management and adaptation of complex and meaningful digital models require the introduction of metamodels, especially when it comes to schema extensions.

It is clear that the digital transformation is changing the traditional concept of documentation and will lead to new challenges for long-term archiving due to increasing virtualization, continuous updating over long periods of time and the embedding of semantically enriched complex 3D models, as is already partly the case in PDF/E-1 (ISO, 2022). Chen, Chang and Lin (2016) even consider so-called dynamic BIM, which includes dynamic data that can be generated, for example, by monitoring the facility's environment or usage. Beach et al. (2017) discuss issues such as data ownership, data security/privacy, and reluctance to share data, and go so far as to consider dynamic federation of distributed BIM models using an overlay technique. The work of Borghoff and Schlichter (1996) provides an early solution with wrappers for dynamic federation in an architectural framework. Another critical case is the derivation of virtual models or documents based on complex numerical algorithms that require additional semantics of the underlying geometry, such as the derivation of 2D drawings for specific purposes or the generation of finite element meshes based on architectural models.

1.3 Research Methodology and Gaps – what is missing?

The research methodology underlying this work is a thorough and comprehensive examination of existing format repositories and relevant projects worldwide. Using an established reference model, we identify research needs and develop a solution to an important aspect of the recognized problem of digital long-term archiving of BIM documents and datasets. As part of a systematic approach to an advanced representation information repository, we develop a metadata model as the centerpiece of the solution, called BIM Core.

We have seen sophisticated solutions for the long-term usability of digital documents in the technical and scientific fields, but they do not address construction, or only to a limited extent; see also Akbari et al. (2024). The implemented repositories or registers lack depth of representation and contextual information or construction-specific content; in particular, there is a lack of identification and description of significant properties, and the current approaches do not take into account the new challenges of BIM and digital transformation described above. In addition, some repository projects have been canceled or stopped after the prototype phase. The lack of commitments or guarantees about the durability of services creates risks that are difficult to assess. We discuss these and other gaps in the next section.

2. RELATED WORK

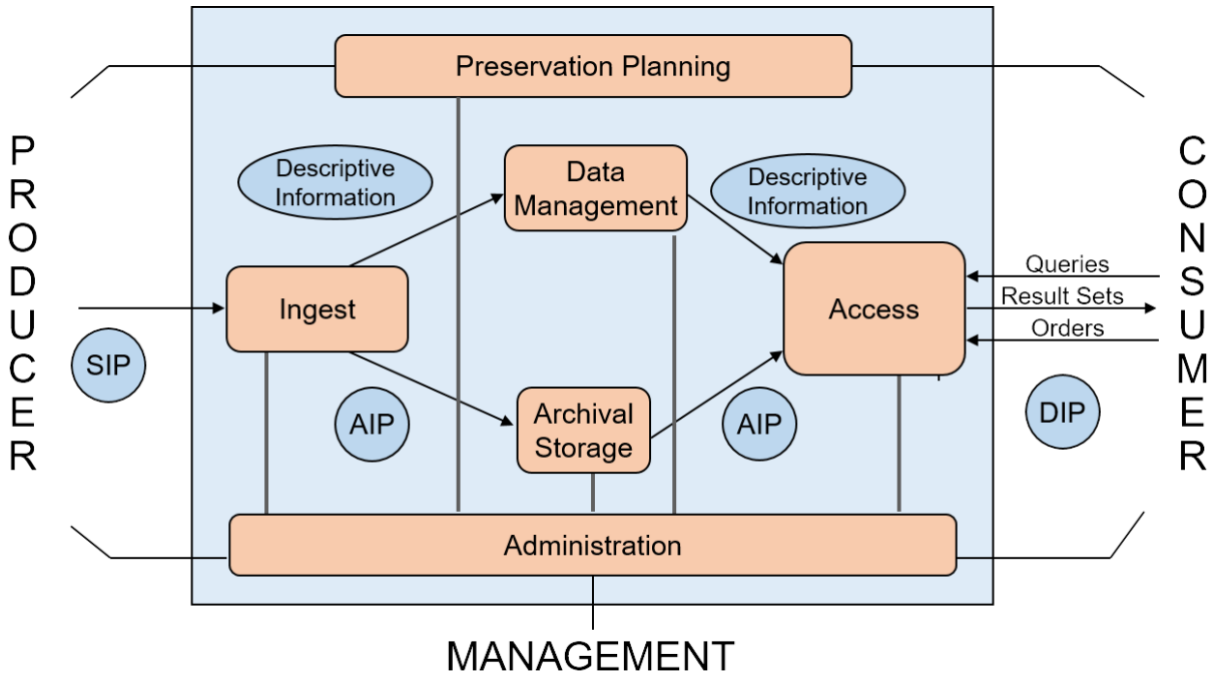

Figure 2: Functional and data entities of the ISO reference model OAIS.

For about a quarter of a century, archives, libraries, scientific research institutions with large and valuable data collections, but also the industry as well as standardization organizations have been dealing with the challenges of digital long-term archiving. The broadly discussed and accepted ISO reference model Open Archival Information System (OAIS) is a milestone for the preservation community (OAIS, 2023), see also Figure 2.

OAIS describes the essential components of an archiving system for long-term information preservation at the conceptual level. It considers data as information packages that contain metadata (e.g., about the data format) in addition to the content. The packages submitted by the producer for archiving are called submission information packages (SIP). They are converted into archive information packages (AIP) upon submission. These contain a variety of information, including the history of previous processing steps as well as necessary migration data and other attributes required for integrity and authenticity. Searching and access by users also takes place via dedicated information packages, the so-called dissemination information packages (DIP).

Key terms include representation information, which includes all the information needed to interpret a bit sequence to convey a meaningful perception to a particular community (Farghaly et al., 2022); see also Bohne et al. (2011). Format descriptions are an important form of representation information and therefore, format registries play an important role in the community.

A further concept of OAIS is *context information*, which contributes to a deeper understanding of digital content; e.g., user manuals or specifications of measuring devices.

Additionally, the digital preservation community has introduced and broadly discussed the concept of *significant properties*, which is not explicitly described in OAIS, but we think that the following definition is suitable in our context: The characteristics of an information object that must be maintained over time to ensure its continued access, use, and meaning, and its capacity to be accepted as evidence of what it purports to record. For details, refer to Van Veenendaal et al. (2018).

Taking into account the provision of relevant content or conceptual and technical reusability, we have considered the following format registries:

- *PRONOM* of the National Archive UK (PRONOM, 2024). This technical registry is an online information system about file formats, their supporting software products, and migration paths. The system contains a large number of formats that are uniquely identified. However, building-specific formats are not fully or partially described. It also lacks elements that we address in our proposed BIM Core.
- The *Global Digital Format Registry* (GDFR)/*Unified Digital Formats Registry* (UDFR) (Goethals, 2010). The GDFR project has been discontinued, and the UDFR service, which was intended to integrate GDFR and PRONOM, was discontinued in 2016. However, the precise conceptual basis of the GDFR is still useful for defining or refining BIM Core elements.
- The *Library of Congress* (LOC), USA, offerings called Sustainability of Digital Formats - Planning for Library of Congress Collections, which also contain evaluation criteria for the sustainability of formats (LOC, 2024). The Web sites associated with this ongoing initiative provide detailed descriptions of file formats in several categories, including design and 3D formats. There is even clear information about the EXPRESS data modeling language. However, formats relevant to the construction sector are still missing, especially with regard to the emerging data management techniques associated with BIM. The work of the LOC is aimed at the needs of the library community, but we believe that the concepts are also suitable for the BIM Core, particularly in terms of the level of detail provided.
- The *Catalog of Archival File Formats* with evaluation systematics for archives from the Swiss institution KOST (Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen) (KOST, 2024). This catalog is online and up-to-date. However, it only lists some building-specific formats and does not include elements that we consider important to cover the use cases we have identified for the BIM Core. The catalog is focused on archives and similar institutions.
- The *File Format Database* from Sharpened Productions, USA (FileInfo, 2024). This database is online and up-to-date. It provides basic information about file formats and, in some cases, a list of programs that can open or convert a file. However, building-specific formats are not complete, and elements we propose for the BIM Core are not included.

Two developments considered the concept of *representation information* explicitly:

- The *Registry/Repository of Representation Information for Engineering* (RRoRIfE, 2024) was intended as a tool for CAD/CAM/CAE preservation planning. To our knowledge, no system is or has been in use in the building sector. However, the discussions and concepts are helpful in the development of the BIM Core, as issues such as significant properties and OAIS representation information are addressed (Patel, Ball and Ding, 2009).
- The *Digital Curation Centre Representation Information Repository*, which was intended for a broader scope (CASPAR, 2024). The project has been discontinued.

The Massachusetts Institute of Technology (MIT) has also specifically addressed the preservation of CAD models and related documents in the field of architecture for scientific purposes, see for example the FACADE project (future-proofing architectural computer-aided design) (MIT FACADE, 2009).

The best-known tool in the field of memory organizations for identifying and validating different file formats is JHOVE (JSTOR/Harvard Object Validation Environment) (JHOVE, 2024). Another means with a long tradition for recognizing a variety of file or file system formats is the UNIX program file (Open Group, 2024). DPC (Digital Preservation Coalition, UK) has been dealing with preserving CAD again and issued a short guide in 2021, which contains concise descriptions of a selection of formats commonly used in the construction sector (DPC, 2024).

The goal of the European research project *DURAble ARchitectural Knowledge* (DuraArK) was to develop methods and tools for the digital curation and preservation of 3D building data, metadata, associated knowledge and web data (European Commission, 2016). The work on building data focused on BIM and scanned point clouds, i.e. the open standardized file format ASTM E2807-11 (ASTM, 2024). A technical metadata schema was developed for each format (ifcm and e57m). The team also designed a descriptive metadata schema that includes a section for the physical asset and one for the digital object (buidm).

MonArch is a digital archive for architectural monuments (Freitag and Schlieder, 2009). The MonArch software combines the digital structural documentation of buildings with semantic descriptions and allows users to document historical buildings and archaeological sites as well as existing buildings in urban areas. It includes a metadata repository that can be used for spatially capturing, semantically tagging, managing, and storing digital documents (Stenzer et al., 2011). The latest version of the system, MonArch 3, was released in February 2024 (MonArch, 2024). A new concept for describing the complexity of historical buildings, which emphasizes both their physical and topological relationships through time and space, is introduced by Ronzino et al (2016). Koehl et al. (2015) propose an example of a highly complex multimedia description of a specific historical building.

The ongoing *LOTAR International project* initiated by the aerospace industry is a noteworthy and probably the most advanced approach in the technical field (LOTAR, 2024). The goal of LOTAR is to develop, test, publish and maintain standards for long-term archiving of digital product and technical data, such as 3D CAD/CAM and PDM data. These standards will define auditable archiving, management, and retrieval processes that preserve engineering intent throughout the product lifecycle. The results are based on the ISO 14721, Open Archival Information System (OAIS) Reference Model. Therefore, LOTAR's work also includes the elaboration and refinement of concepts related to information packages (i.e., SIP, AIP, DIP). In addition, the LOTAR consortium has established a "Metadata for Archival Package" working group whose goal is to develop, publish, and maintain a standard for metadata for Archival Packages in a neutral form that can be read and reused regardless of changes in the IT application environment originally used to create it. The planned NAS (National Aerospace Standards)/EN (European Norm) 9300 Part 021 standard "Long Term Archiving and Retrieval of Meta data for Archival Packages" is currently (July, 2024) TBC (to be confirmed). LOTAR also addresses standards-based mechanisms for archiving and retrieving engineering analysis and simulation information, such as finite element model data and internal load results needed to analyze structural repairs, or finite element analysis to support development studies of a new product design.

Despite the differences with the construction industry, we also see similarities, such as the complexity of the products, the large number of suppliers, the relatively long service life, strict legal requirements for documentation, and the transition from document-based to model-based information management. It is therefore useful to take a closer look at LOTAR's approaches and discussions, e.g. regarding the structure, content and metadata of information packages, the importance of neutral formats, the preservation of design intent, the definition of significant properties (referred to as validation properties), and the capture and preservation of knowledge. For us,

the LOTAR project is also a confirmation that OAIS is a suitable basis for implementing long-term archiving solutions in the technically demanding area of (BIM-based) knowledge management.

Wang and Meng (2019) provide a literature review of the transition from traditional information technology for knowledge management to BIM-supported knowledge management. They present a conceptual model for BIM-supported knowledge management that highlights the intricate web of potential factors and their interrelationships within the BIM ecosystem.

3. FIRST RESULTS

Borghoff, Pfeiffer and Rödig (2022) present a holistic and standard-based concept for the long-term usability of datasets and documents common to building authorities and other stakeholders that are responsible for large digital assets. Here, we refine an important conceptual element of OAIS, namely the Archival Information Package (AIP). In particular, we focus on the basic concept of data objects (bit sequences) and their interpretation by *representation information*. We describe the concept of *representation information* and present a design for the development of a repository for this type of information, in particular for the storage of BIM documentation. An AIP and its components are shown in Figure 3.

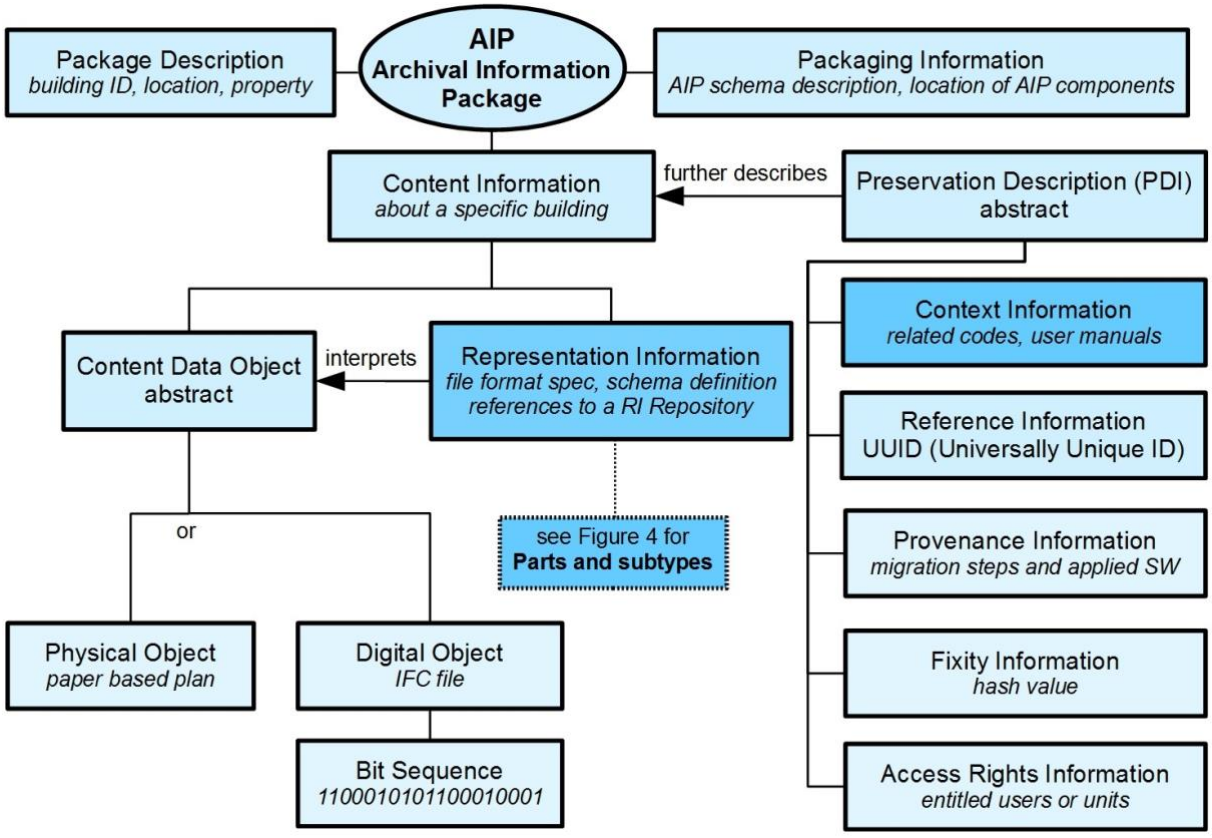

Figure 3: OAIS's data entity Archival Information Package (AIP) and its components.

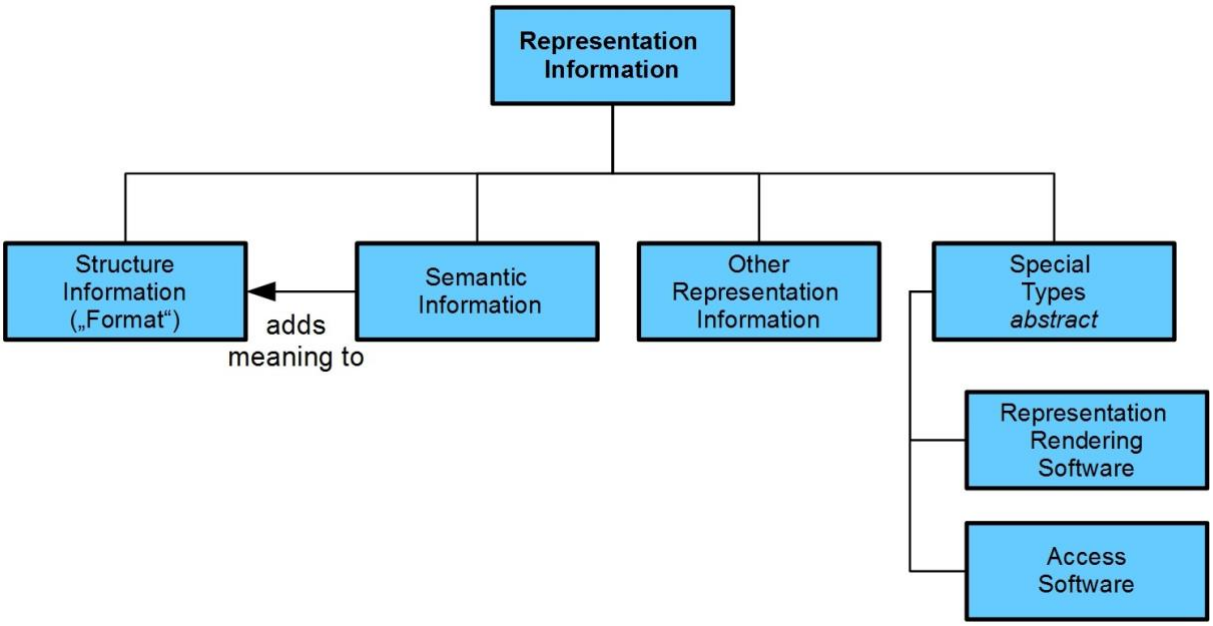

Figure 4: Parts and subtypes of AIP's component Representation Information.

Representation information as a key concept of OAIS is responsible for giving the stored bit sequences meaning for the designated communities.

As Figure 4 depicts, the standard distinguishes between three subtypes of representation information:

- *structure information* should describe the elementary structure of the bit sequences so that all subsequences can be identified and assigned to primitive data types and data structures and further on to higher level concepts; e.g., four subsequences with the length of four bytes in big endian-order each may represent four numbers, which in turn represent the coordinates of the starting and ending point of a 2D-line;
- *semantic information* should provide additional meaning to all the elements described by the structural information, e.g., a set of lines in a defined order and coordinate system may represent a floor plan of a building at a defined scale;
- *other representation information*, which should complement the representation information for the most complete understanding of a data object. OAIS lists these examples: software, algorithms, encryption, and written instructions. In the case of our floor plan, an algorithm could incorporate knowledge of what is supposed to happen when one line is moved.

In addition, OAIS defines two special types of representation information:

- *representation rendering software* must be able to display digitally coded representation information in an understandable form;
- *access software* presents some or all of the information represented by a digital object to humans or systems.

4. THE REPOSITORY

Our research and discussions to date have led to the identification of use cases and, building on these, to an initial definition of conceptual content elements for a representation information repository. Examples in the specific BIM context are used to explain the concept and illustrate how such a repository could be applied.

4.1 Use Cases

We have found that it is desirable to address the following five major roles and use cases:

- *Data and document producer* (SIPs as output) concerned with long-term preservation for legal, economic or cultural purposes.
- *Construction experts as data and document consumer* (DIPs as input) for various levels of reuse where some cases require a high degree of precision and reliability, e.g., nuclear decommissioning or the restoration of culturally valuable buildings.
- *The management and administration of archives* for the evaluation of the quality of formats, semantics and other representation information in support of preservation planning as well as day-to-day tasks such as the construction of AIPs. The focus is on the permanent preservation of information independent of the life cycle of a construction. In particular, this use case requires the evaluation of so-called lightweight formats, as already presented and discussed in Patel, Ball and Ding (2009). This is necessary to minimize the risk of unintentional information loss in the long run.
- *Computer experts* who want to evaluate the design, reuse existing code or implement preservation tools such as emulators or converters for format migrations.
- *Computer and other technical oriented historians*, as electronic data processing has a long tradition, especially in civil and mechanical engineering.

4.2 The BIM Core

In analogy to the *Dublin Core* (Dublin Core, 2024), which contains metadata elements for describing Internet documents, we call our 23 content elements the **BIM Core**. It contains a set of well-structured metadata elements for describing BIM documentation and all types of related data and documents.

To cover the listed use cases, we propose the following *BIM Core content elements*, divided into the three categories: (i) structural and semantic features, (ii) file format handling tools, and (iii) contexts.

4.2.1 Structural and semantic features

We combine structural and semantic aspects, as file formats can also contain semantic information in the sense of OAIS.

Some formats used in the building sector can even be compared to semantic databases, as they explicitly represent domain-specific entities and relationships. Examples include the IFC, as in ISO 16739-1:2024 (ISO, 2024), or multi-model/information container formats, as in ISO 21597-1:2020 (ISO, 2020), which are intended to establish semantic interoperability between data with different formats and domain affiliations.

Therefore, we need as content elements:

1. **name, identifier and classifications** of a format;
2. **version of a format or reference** to other versions;
3. **references** to already existing repositories or registries;
4. **description of forward and backward compatibility** issues;
5. **facts concerning the openness** of a format, especially the availability of specifications;
6. **description of rights and intellectual property** to the formats and, if available, tools that allow reengineering;
7. **description of syntax** (formal structure) **and semantics** (meaning of the individual elements and their relationships) of a format;
8. **encoding of fundamental format elements** such as numbers, texts, colors;
9. **description of compression, data reduction and encryption** of elements or whole files;
10. **references to external standards or documents** used to specify the format or parts of it, e.g., from the following sources: ECMA, ICC, IEEE, ISO, ITU, JPEG, W3C;

11. **self-documentation capabilities** of a format, e. g., the ability to include metadata;
12. **resolvability of external references** (materialization);
13. **description of dependencies** of a format on software and hardware to access or process the content. This is especially the case when devices, e.g., special printers, are controlled;
14. **description of options for tailoring** a format to specific purposes or contexts;
15. **description of background and fundamentals**; associated descriptions or sources may provide *other representation information* as defined by OAIS.

4.2.2 File format handling tools

As mentioned above, some formats may not be obsolete themselves, but the software or settings used to create them may be. This can cause significant problems in interpreting specific content information created during retro-digitization. File format handling tools support the interpretation and quality assurance of content information, as well as the creation and long-term maintenance of AIPs.

It is clear that the processes defined in OAIS need to be automated due to the size of the digital collections. To this end, we introduce the following three additional content elements:

16. **description of software tools for the recognition** of formats;
17. **description of software tools for the verification** of formats and, if applicable, tailored versions against their specification;
18. **description of software tools for inspection and (simplified) presentation of the content** (see *access software* as a special type of representation information in Figure 4).

4.2.3 Contexts

This category should provide information that is necessary for a deeper and a long-term understanding of the digital content and for preservation planning, e.g., for assessing file format migrations, in particular, with regard to *significant properties*.

Although *context information* is a separate conceptual element in OAIS/AIP (see Figure 3), we propose to integrate such information into a representation information repository because of the complexity of content information and the problems in distinguishing between *representation information* and *context information*. Additionally, we suggest to assign the technical aspects of *provenance information* to *context information*, for example, specification documents of instruments like sensors as an origin or source of generated data. Indeed, OAIS states that provenance could be viewed as a special type of *context information*.

However, we have to consider that parts of the *context information* may depend on organizational, local or project-specific circumstances, which in addition requires individualized repositories or an explicit integration into AIPs. For example, project manuals are a good source for describing the context of a construction project since they typically contain applied standards and guidelines as well as specifications for reporting and project documentation. The developers of the format IFC have also recognized the importance of *context information* and therefore introduced an entity called IfcContext which is the generalization of a project context in which objects, type objects, property sets, and properties are defined. IFC is an example for a format that can incorporate several components of an AIP in a standardized way. To make the list complete, we propose five more content elements:

19. **description of typical contexts of origin and significant properties** of typical application scenarios to be preserved;
20. **incorporation of building-specific context information**;
21. **description of acceptance and frequency of use** of a format within a specific context;
22. **description or sources of background and fundamentals**;
23. **references to repositories containing use cases or even significant properties**.

A starting point for a shared, community-owned database with explicit knowledge on significant properties as in Content Element #23 consider van Veenendaal et al. (2018). In other cases, use cases can help to identify and describe significant properties. For example, buildingSMART International's Use Case Management Service (UCMS) enables the capture, specification and sharing of best practices and makes them available to the entire built asset industry (buildingSMART, 2024). This is indeed necessary, as Hu et al. (2022) state that in the past, the data and even some of the knowledge stored in the BIM database were mostly useful only for a specific project, but not for the general purpose.

Finally, it should be noted that the conceptual content elements of the repository will largely consist of digital objects and will therefore also require representation information for their long-term understanding. Representation rendering software, as a special type of representation information, should be able to present it in a way that is understandable in the long term. Of course, access software, as another special type, requires additional considerations.

4.3 Application of the Representation Information Repository in a BIM Context

In this section, we explain how the proposed representation information repository can be used to implement the two OAIS concepts of representation information and context information, as well as the concept of significant properties, which is not explicitly modeled in OAIS. We focus here on the specifics of the construction sector and BIM; therefore, we omit details found in general purpose file format repositories, such as IDs, rights and intellectual property issues, or metadata for managing such a repository. Even medium-sized construction projects require a large number of digital documents and datasets (five hundred or more) that reflect the information and documentation needs of stakeholders such as owners, architects, engineers, contractors, project controllers, and public administrators. Each of them uses their own tools with specific formats, schemas and configurations. We have found that common and widely used tools and formats are in use, especially for word processing and spreadsheets, but in practice, domain-specific and sometimes very complex formats from different vendors dominate. Since buildings generally have a long life span, digital assets also contain obsolete formats. The associated heterogeneity leads to problems with data exchange, data consistency and correct interpretation, especially in the long term. BIM is a widely accepted approach to minimizing these shortcomings, with the semantically enriched and standardized IFC format, including evolving co-standards and regulations, playing a key role; see Farghaly et al. (2022) for a recent application. The following example illustrates how the proposed layout of a repository for representation information can be applied. We start with an IFC file that is to be kept for the operation of a building and for reuse. Since IFC is now a widely used format, information about it can be found in PRONOM (PRONOM, 2024) or on the LOC web pages (LOC, 2024).

Because IFC covers various disciplines and aspects at multiple levels of detail and development, it is common to tailor the format to specific purposes by extending or limiting the standard schema, e.g., to meet information needs for verifying and documenting compliance with fire codes.

Now contextual information comes into play as another concept of OAIS in our use case. To fully understand the content and logic represented by such a customized format, we need to know exactly what the underlying rules are and how their content can be reliably accessed—even decades later. Therefore, we also need *representation information* (OAIS: *representation rendering software* as a special kind of *representation information*) when the referenced context documents are digital objects. Fortunately, these types of documents are generally encoded in a relatively simple, common, and standardized format. It would be good practice to keep such general-purpose documents that can provide *context information* in a long-term managed repository. However, *context information* or *representation information*, including written instructions or algorithms (in OAIS: *other representation information* as a subtype of *representation information*) specific to a project or document must be integrated into an AIP, the structure of which is shown in Figure 3.

Since IFC is a rather complex format, it is desirable to have formalized and standardized tools available for customization. Such a tool, called *Information Delivery Service* (IDS), is under development, and the underlying file format is another candidate for a representation information repository. An IDS editor is an example of *access software* as another specific type of *representation information* of the OAIS (Content Element #12). A typical use case for IDS is, for example, the adaptation of IFC building models to the requirements of fire protection (Content Element #19). In this use case, the adaptation rules reflect the significant properties of such an information object, since these rules describe fairly directly the characteristics that must be maintained over time (Content Element

#19). With the advent of digital twins and BIM, dynamic data is becoming increasingly important for assessing the quality of a construction, including its safety status. Dynamic data, provided continuously or periodically by sensors to monitor energy consumption or the structural condition of buildings, for example, requires special consideration in process and data models. Various approaches have been developed to integrate monitoring data and sensor information into BIM/IFC, such as using graph database systems for data and IFC elements, or explicitly modeling sensor networks using extended IFC schemas. Even ontologies for sensor networks are available. In order to ensure the comprehensibility of data, especially when long-term monitoring is implemented, information on formats and instruments is required. Therefore, there is a reference to Content Element #1, which then provides the information stored in Content Elements #2 through #15. For example, in the case of long-term monitoring, information on forward and backward compatibility (Content Element #4) may be required to assess whether the recorded data can be compared or migrated if formats have changed.

If standard tools and monitoring techniques are tailored for building specific purposes, the appropriate information should be included in Content Element #20. Again, especially if a more complex format or schema is used, it is useful for the data producer to find information about tools for verifying the data they generate or receive for further processing (Content Elements #16, #17, and #18). Please note that dynamic data management and metadata support, such as updating and versioning of AIPs, is outside the scope of a representation information repository. However, identifying and defining significant properties can be a challenging task, especially when complex architectural or technical drawings and models for structural analysis, fluid mechanics or similar topics need to be stored for later reuse.

5. CONCLUSION

This paper illustrates the benefits and challenges of long-term usability of digital documents and datasets in a complex application domain, and then presents an approach to a key problem: preserving the meaning of digital objects with the support of a repository that provides *representation information* and *context information*. We recognize that the design, data collection, and long-term operation and maintenance of such a repository is a challenging task that can only be accomplished in cooperation with all relevant stakeholder, including memory organizations, building administrations and the construction sector.

During our research, we found that a better theoretical foundation would be helpful, especially with respect to first, representation information including their subtypes, and second, the inclusion of *significant properties* and the two OAIS conceptual elements of context and provenance information to give a repository a clear design. Another arduous task is to find and use all the knowledge sources scattered all over the world to populate such a repository with useful information.

We have introduced the BIM Core, which will help with the fine structuring needed during the capture process. The goal of the BIM Core is to be a simple metadata model that can be adopted by a wide range of communities in an effort to improve semantic interoperability. In this sense, we believe that our approach will also be very useful for other scientific and technical disciplines.

ACKNOWLEDGEMENT

Our work on building authorities is in part funded by the Federal Ministry of the Interior, Building and Community (BMI) Grant Ref: SWD 10.08.18.7-15.36 under the Future Building programme.

We would like to thank our partners from Vintage Computing Lab (VCL, 2024) and Cray-Cyber.org (Cray-Cyber, 2024), who support the operation of the datArena as an institution of the University of the Bundeswehr Munich. The datArena has an extraordinary pool of hardware, software and documentation and is dedicated to the preservation of the digital cultural heritage (datArena, 2024).

Special thanks also go to the Bavarian Building Administration, the General Directorate of the Bavarian State Archives, Munich, Germany, for presenting their challenges and needs regarding the long-term usability of digital building information, and to the software and consulting companies novaCapta and Thinkproject for the intensive discussion on long-term data management in the construction sector.

REFERENCES

Akbari, S., Sheikhkhoshkar, M., Pour Rahimian, F., Bril El Haouzi, H., Najafi, M., & Talebi, S. (2024). Sustainability and building information modelling: Integration, research gaps, and future directions. Automation in Construction, 163. doi:10.1016/j.autcon.2024.105420

Alshammari, K., Beach, T. H., & Rezgui, Y. (2021). Cybersecurity for digital twins in the built environment: Current research and future directions. Journal of Information Technology in Construction, 26, pp. 159-173. doi:10.36680/j.itcon.2021.010

ASTM (2024). Retrieved from https://www.astm.org/get-involved/technical-committees/committee-e57

Barton, C. I. (2006). Elements of a Good Document Retention Policy. LexisNexis, Applied Discovery White Paper. Retrieved from https://www.lexisnexis.ca/

Beach, T., Petri, I., Rezgui, Y., & Rana, O. (2017). Management of Collaborative BIM Data by Federating Distributed BIM Models. Journal of Computing in Civil Engineering, 31(4). doi:10.1061/(ASCE) CP.1943-5487.000065

Beinert, T., Lang, S., Schoger, A., Borghoff, U. M., Hagel, H., Minkus, M., & Rödig, P. (2008). Development of Organisational and Business Models for the Long-Term Preservation of Digital Objects. iPres Conf., London, UK. Retrieved from https://phaidra.univie.ac.at/detail/o:294056

Bohne, T., Rönnau, S., & Borghoff, U. M. (2011). Efficient keyword extraction for meaningful document perception. ACM DocEng Conf., Mountain View, CA, USA, pp. 185-194. doi:10.1145/2034691.2034732

Borghoff, U. M., & Schlichter, J. H. (1996). On Combining the Knowledge of Heterogeneous Information Repositories. J. Univers. Comput. Sci., 2(7), pp. 515-532. doi:10.3217/jucs-002-07-0515

Borghoff, U. M., Pfeiffer, E., & Rödig, P. (2022). Long-Term Lifecycle-Related Management of Digital Building Documents: Towards a Holistic and Standard-based Concept for a Technical and Organizational Solution in Building Authorities. ACM DocEng Conf., San Jose, CA, USA, pp. 6:1-6:10.

doi:10.1145/3558100.3563842

Borghoff, U. M., Rödig, P., Schmitz, L., & Scheffczyk, J. (2006). Long-term preservation of digital documents - principles and practices. Springer. doi:10.1007/978-3-540-33640-2

buildingSMART (2024). Retrieved from https://ucm.buildingsmart.org/

CASPAR (2024). Retrieved from https://www.dcc.ac.uk/guidance/briefing-papers/technology-watch-papers/

Chen, H.-M., Chang, K.-C., & Lin, T.-H. (2016). A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs. Automation in Construction, 71, pp. 34-48. doi:10.1016/j.autcon.2016.03.002

Chen, X., Chang-Richards, A. Y., Pelosi, A., Jia, Y., Shen, X., Siddiqui, M., & Yang, N. (2022). Implementation of technologies in the construction industry: A systematic review. Engineering, Construction and Architectural Management, 29(8), pp. 3181-3209. doi:10.1108/ECAM-02-2021-0172

Choy, S. C., Crofts, N., Fisher, R., Choh, N. L., Nickel, S., Oury, C., & Ślaska, K. (2016). UNESCO/PERSIST Content Task Force. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000244280

Cray-Cyber (2024). Retrieved from https://cray-cyber.org

Cursi, S., Martinelli, L., Paraciani, N., Calcerano, F., & Gigliarelli, E. (2022). Linking external knowledge to heritage BIM. Automation in Construction, 141. doi:10.1016/j.autcon.2022.104444

datArena (2024). Retrieved from https://www.unibw.de/inf2/forschung/forschungsthemen/datarena

Dobratz, S., Rödig, P., Borghoff, U. M., Rätzke, B., & Schoger, A. (2010). The Use of Quality Management Standards in Trustworthy Digital Archives. International Journal of Digital Curation, 5(1), pp. 46-63. doi:10.2218/ijdc.v5i1.143

DPC (2024). Retrieved from https://www.dpconline.org/docs/technology-watch-reports/2480-preserving-cad/file

Dublin Core (2024). Retrieved from https://www.dublincore.org/

European Commission (2016). Retrieved from https://cordis.europa.eu/project/id/600908

Farghaly, K., Soman, S. K., Collinge, W., Mosleh, M. H., Manu, P., & Cheung, C. M. (2022). Construction safety ontology development and alignment with industry foundation classes (IFC). Journal of Information Technology in Construction, 27, pp. 94-108. doi:10.36680/j.itcon.2022.005

FileInfo (2024). Retrieved from https://fileinfo.com

Forum. (2024, January 8). Why a digital and AI-first approach is the fastest path to net-zero buildings. Retrieved from The World Economic Forum: https://www.weforum.org/

Freitag, B., & Schlieder, C. (2009). MonArch - Digital Archives for Monumental Buildings. Künstliche Intelligenz, 23(4), pp. 30-35.

Ge, X. J., Livesey, P., Wang, J., Huang, S., He, X., & Zhang, C. (2017). Deconstruction waste management through 3d reconstruction and BIM: A case study. Visualization in engineering, 5(13), pp. 1-15. doi:10.1186/s40327-017-0050-5

GI (2014). Die Grand Challenges der Informatik. Gesellschaft für Informatik. Retrieved from https://gi.de/grand-challenges

Goethals, A. (2010). The Unified Digital Formats Registry. Information Standards Quaterly, Special Issue on Digital Preservation, 22(2), pp. 26-29.

Golshan, B., Halevy, A. Y., Mihaila, G. A., & Tan, W.-C. (2017). Data Integration: After the Teenage Years. ACM PODS, Chicago, IL, USA, pp. 101-106. doi:10.1145/3034786.3056124

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data Integration: The Teenage Years. ACM VLDB Conf., Seoul, Korea, pp. 9-16. doi:10.5555/1182635.1164130

Hodge, G., & Anderson, N. (2007). Formats for digital preservation: A review of alternatives and issues. Information Services & Use, 27(1-2), pp. 45-63. doi:10.3233/ISU-2007-271-204

Hu, Z.-Z., Leng, S., Lin, J.-R., Li, S.-W., & Xiao, Y.-Q. (2022). Knowledge Extraction and Discovery Based on BIM: A Critical Review and Future Directions. Archives of Computational Methods in Engineering, 29, pp. 335-356. doi:10.1007/s11831-021-09576-9

ISO (2020). Retrieved from https://www.iso.org/standard/74389.html

ISO (2022). Retrieved from https://www.iso.org/standard/42274.html

ISO (2024). Retrieved from https://www.iso.org/standard/84123.html

JHOVE (2024). Retrieved from https://jhove.openpreservation.org/documentation

Koehl, M., Viale, A., & Reeb, S. (2015). A Historical Timber Frame Model for Diagnosis and Documentation before Building Restoration. International Journal of 3-D Information Modeling, 4(4), pp. 34-63. doi:10.4018/IJ3DIM.2015100103

KOST (2024). Retrieved from https://kost-ceco.ch/cms/kad_main_de.html

Leventhal, A., Thompson, J., Anderson, A., Schubert, S., & Altenbach, A. (2021). Design Records Appraisal Tool. The American Archivist, 84(2), pp. 320-354. doi:10.17723/0360-9081-84.2.320

Li, Y., Du, Y., Yang, M., Liang, J., Bai, H., Li, R., & Law, A. (2023). A review of the tools and techniques used in the digital preservation of architectural heritage within disaster cycles. Heritage Science, 11(1), pp. 1-20. doi:10.1186/s40494-023-01035-x

LOC (2024). Retrieved from https://www.loc.gov/preservation/digital/formats

LOTAR (2024). Retrieved from https://lotar-international.org

Lou, J., Lu, W., & Xue, F. (2021). A Review of BIM Data Exchange Method in BIM Collaboration. CRIOCM Symposium. Springer, Singapore, pp. 1329–1338. doi:10.1007/978-981-16-3587-8_90

Madubuike, O. C., Anumba, C. J., & Khallaf, R. (2022). A review of digital twin applications in construction. Journal of Information Technology in Construction, 27, pp. 145-172. doi:10.36680/j.itcon.2022.008

MIT FACADE (2009). Retrieved from https://www.vai.be/volumes/general/FACADEFinalReport.pdf

MonArch (2024). Retrieved from https://openmonarch.org/2024/02/monarch-version-2024-02/

OAIS (2023). ISO 14721:2012 Space data and information transfer systems — Open archival information system (OAIS) — Reference model; reviewed and confirmed in 2023. Retrieved from https://www.iso.org/standard/57284.html

Open Group (2024). Retrieved from https://pubs.opengroup.org/onlinepubs/9699919799/utilities/file.html

Panah, R. S., & Kioumarsi, M. (2021). Application of building information modelling (BIM) in the health monitoring and maintenance process: A systematic review. Sensors, 21(3). doi:10.3390/s21030837

Patel, M., Ball, A., & Ding, L. (2009). Strategies for the Curation of CAD Engineering Models. International Journal of Digital Curation, 4(1), pp. 84-97. doi:10.2218/ijdc.v4i1.80

Pocobelli, D. P., Boehm, J., Bryan, P., Still, J., & Grau-Bové, J. (2018). BIM for heritage science: A review. Heritage Science, 30, pp. 1-15. doi:https://doi.org/10.1186/s40494-018-0191-4

PRONOM (2024). Retrieved from https://www.nationalarchives.gov.uk/aboutapps/pronom/

Ronzino, P., Niccolucci, F., Felicetti, A., & Doerr, M. (2016). CRMba a CRM extension for the documentation of standing buildings. International Journal on Digital Libraries, 17(1), pp. 71-78. doi:10.1007/s00799-015-0160-4

RRoRIfE (2024). Retrieved from https://rorife.sourceforge.net/

Stenzer, A., Woller, C., & Freitag, B. (2011). MonArch: Digital archives for cultural heritage. ACM iiWAS Conf., Ho Chi Minh City, Vietnam, pp. 144-151. doi:10.1145/2095536.2095562

Toyin, J. O., Azhar, S., Sattineni, A., & Fasoyinu, A. A. (2024). Investigating the Influence of ICT Application in Construction Jobsites: A Systematic Review and Bibliometric Analysis. Journal of Information Technology in Construction, 29, pp. 444-479. doi:10.36680/j.itcon.2024.021

Van Veenendaal, R., Lucker, P., & Sijtsma, C. (2018). Significant Significant Properties. Retrieved from https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf

VCL (2024). Retrieved from https://www.vclab.de

Wang, H., & Meng, X. (2019). Transformation from IT-based knowledge management into BIM-supported knowledge management: A literature review. Expert Systems with Applications, 121, pp. 170-187. doi:10.1016/j.eswa.2018.12.017

Waqar, A., & Ahmed, W. (2023). Reimagining construction safety: Unveiling the impact of building information modeling (BIM) implementation. Safety in Extreme Environments, 5(4), pp. 265-280. doi:10.1007/s42797-023-00086-4

Weili, L., Khan, H., Khan, I., & Han, L. (2022). The impact of information and communication technology, financial development, and energy consumption on carbon dioxide emission: Evidence from the Belt and Road countries. Environmental Science and Pollution Research, 29, pp. 27703–27718. doi:10.1007/s11356-021-18448-5

Zahedi, A., Abualdenien, J., Petzold, F., & Borrmann, A. (2022). BIM-based design decisions documentation using design episodes, explanation tags, and constraints. Journal of Information Technology in Construction, 27, pp. 756-780. doi:10.36680/j.itcon.2022.037