HAL
archives-ouvertes.fr

Toward a Reference Architecture for Archival Systems

Raphael Barbau, Joshua Lubell, Sudarsan Rachuri, Sebti Foufou

▶ **To cite this version:**

Raphael Barbau, Joshua Lubell, Sudarsan Rachuri, Sebti Foufou. Toward a Reference Architecture for Archival Systems. Alain Bernard; Louis Rivest; Debasish Dutta. 10th Product Lifecycle Management for Society (PLM), Jul 2013, Nantes, France. Springer, IFIP Advances in Information and Communication Technology, AICT-409, pp.68-77, 2013, Product Lifecycle Management for Society. <10.1007/978-3-642-41501-2_8>. <hal-01461922>

HAL Id: hal-01461922
https://hal.inria.fr/hal-01461922

Submitted on 8 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Toward a reference architecture for archival systems

Raphael Barbau[1,2], Joshua Lubell[2], Sudarsan Rachuri[2], and Sebti Foufou[3,1]

[1] Le2i, Université de Bourgogne,BP 47870, 21078 Dijon, France
[2] National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[3] CSE Department, CENG, Qatar University, Doha, Qatar

Abstract. Long-term preservation of product data is imperative for many organizations. A product data archive should be designed to ensure information accessibility and understanding over time. Approaches such as the Open Archival Information System (OAIS) Reference Model and the Audit and Certification of Trustworthy Digital Repositories (ACTDR) provide a framework for conceptually describing and evaluating archives. These approaches are generic and do not focus on particular contexts or content types. Enterprise architecture provides a way to describe systems in their potentially complex environments.
This paper proposes a holistic approach to formally describe the architecture and the environment of archival systems. This approach relies on the formal representation of the preservation terminology, including OAIS concepts, using the Department of Defense Architecture Framework (DoDAF). The approach covers the various interactions of other business functions with the archive, and the information models necessary to ensure preservation and accessibility. This approach is a step toward a reference architecture for the formal description of archival systems.

1 Introduction

A large amount of digital information is produced and consumed every day. Although most of this information is for immediate consumption, many organizations have an interest in long-term preservation[1]. The main motivations for preserving information are reusing existing knowledge, or keeping proofs of past events. Besides typical digital data management, specific information and activities are needed to ensure data's long-term preservation and accessibility[2, 3]. These information objects and activities are part of a dedicated entity: the archive. Design of an archive is a key factor in successful preservation, especially when complex information and activities are involved. Product data preservation is a good example of such complexity.

A complex product may be composed of numerous systems and parts. For each part, various product data may be produced from conception to disposal. Product data may be formally represented through large and complex information models. The metadata necessary to organize, interpret, or prove the

authenticity of the information may also be complex. Finally, the interactions between the different product data repositories and the archive may be complex, and may involve many different stakeholders at different product lifecycle stages. One issue of this complexity is that the production and the consumption of preserved information is usually part of business functions not dedicated to preservation. This means that the archive has to be well integrated within the organization.

Some efforts address product data preservation by proposing alternative representations for the content [4, 5]. This paper focuses more on the infrastructure aspect of product data preservation, and more particularly on the computer systems involved in the preservation: the archival systems.

The design of archival systems should include the information, activities, systems, and other concepts needed to carry out the preservation mission. In the most complex situations, a product models archival system may have to communicate with different data sources (e.g. Product Data Management, Enterprise Resource Planning, Maintenance and Repair Operations, etc.). Communications among systems, activities, and information have to be well defined and interrelated in a consistent way. The objective of this paper is to propose a way to formally describe these elements.

The modeling of systems and their environment in the context of an enterprise is addressed by Enterprise Architecture (EA). EA establishes a link between the missions of an organization and their implementations. EA typically supports the description of systems, services, activities, information, and constraints within an organization. In our case, EA can be leveraged to detail how the preservation strategy is implemented.

Different efforts have attempted to determine the common elements that constitute archival systems. The Reference Model for an Open Archival Information System (OAIS RM) [6] is a mature conceptual framework for describing and comparing archives. It defines a common terminology for information preservation, especially from the information and functional perspectives. The OAIS RM has been adopted in various product data preservation efforts[7, 8]. However, its models are generic and conceptual: they are not meant to be directly implemented, but rather to serve as guidance for preservers to develop their own solutions.

This paper presents an approach that combines the concepts and terminology defined in the OAIS RM with those used in EA to allow formal description of archival system architectures. By using a formal description, the preservation concepts are explicitly referred to, which increases the understanding of the design, ensures consistency among the various elements described, and ultimately leads to an implementation of high quality. This approach is a first step towards the definition of a generic reference architecture to guide and constrain the description of archival systems.

This paper is organized as follows. Section 2 presents background information on preservation and enterprise architecture. Section 3 presents the approach for

combining the OAIS RM and EA to enable the description of an archival system. Finally, Section 4 presents our conclusions.

2 Background on digital preservation and enterprise architecture

This section provides background information about the conceptualization, development, and certification of archival systems. It also introduces enterprise architecture, and particularly the Department of Defense Architecture Framework used in our approach.

2.1 Conceptual frameworks and certification of archival systems

Reference Model for an Open Archival Information System (OAIS RM), also known as ISO 14721, proposes a conceptual framework for describing and comparing archives[6]. It defines the terminology related to information preservation, including the types of information required to ensure preservation and accessibility of the content, and the main functions that an archive should support.

The OAIS RM defines the different kinds of information in an archive. This information, composed of content and metadata, is encapsulated in information packages. The Submission Information Package (SIP) refers to what the producer sends to the OAIS. The Archival Information Package (AIP) refers to what the archive stores. The Dissemination Information Package (DIP) refers to what the archive delivers to the consumer. Preservation Description Information (PDI) refers to information added to the content to ensure its preservation. Descriptive Information (DI) is a subset of PDI used to locate the desired information.

The OAIS RM describes the main functions of an archive (see Figure 1). The Ingest function receives the SIPs, and generates AIPs to be sent to Archival Storage, and DI to be sent to Data Management. The Data Management function sends some DI to the Access function when needed, and the Archival Storage function sends the desired AIP to the Access function. Then, the Access function returns a DIP to the consumers. The Preservation Planning function monitors the environment of the OAIS. The Administration function, directed by the management, establishes the overall preservation strategy of the OAIS. Each function is further decomposed into smaller functions in the OAIS RM.

Both information and functions are presented in a conceptual and generic way: they are not tied to a particular domain or implementation method. Actual solutions need to be tailored to the specific preserved content and to the context of the preservation. Also, the OAIS RM does not make the distinction between functions done by humans and functions performed by computers. However, it is unclear how to formally incorporate this terminology within actual archival system designs.

Audit and Certification of Trustworthy Digital Repositories (ACTDR)[9] is a standard for the certification of an OAIS. It addresses organizational aspects that are not considered in the OAIS RM, and it gives more details about what

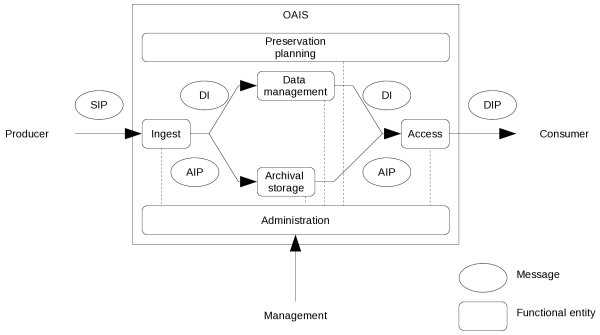

Fig. 1. OAIS Functional Model

is expected from the archive. The certification concerns three different areas: the organizational infrastructure, the digital object management and the infrastructure and security risk management. Each area is composed of requirements and examples of how to demonstrate that the organization meets that requirement.

2.2 Introduction to enterprise architecture and its use for information preservation

Our approach to design archival systems is to rely on enterprise architecture. EA is the discipline of formally describing an enterprise, in particular the systems that compose it. An enterprise can be defined as an organization or a subset of an organization. EA describes how the objectives of an enterprise are realized through systems, services, and activities [10]. EA provides an abstract view that makes it easier to understand how the enterprise works, and how the systems are integrated. The actual description of an enterprise or of one of its parts is called architectural description.

Using EA for representing systems requires two components: a method that provides the steps in the development of the architecture, and the tools to concretely describe this architecture, for example by providing a metamodel. Different Enterprise Architecture Frameworks (EAFs) propose varying approaches to describe enterprises, and sometimes they focus on the aspect they judge the most important. For example, The Open Group Architecture Framework (TOGAF) [11] is an EAF well known for its Architectural Development Method. TOGAF has not incorporated a metamodel until recently. On the other hand, the Ministry of Defence Architecture Framework (MODAF)[12] and the US Department of Defense Architecture Framework (DoDAF)[13] focus on defining a metamodel and a set of views to formally represent architectural descriptions. Other EAF include the Generalized Enterprise Reference Architecture and Methodology (GERAM), developed by the IFIP-IFAC Task Force [14], and included as an Appendix of ISO15704:2000 [15] is a generalized EAF for enterprise integration and business process engineering. GERAM defines all the components required

for use in enterprise engineering. Other well-known reference architectures are the Purdue Enterprise Reference Architecture (PERA)[16], and CIMOSA[17].

Becker et al. presented a reference architecture approach for archives, which emphasized the development process rather than the description of the actual implementations[18]. Becker et al.'s approach was to highlight the different recommendations and standards that need to be considered during the design of the archive. Becker et al. took into account preservation recommendations and information technology standards related to risk management, data quality, or security. However, they did not address the formal description of the archival system architecture.

This paper presents an approach that focuses on the formal description of archival system architectures. Our approach consists of 1) formally representing the preservation terminology to be used in architectural descriptions, and 2) defining a set of views to describe important aspects of an archival system. We rely on DoDAF to provide the core enterprise architecture concepts to which preservation concepts relate. DoDAF provides a metamodel made of generic concepts, which don't include domain-specific concepts such as preservation concepts[13]. A formal description offers several benefits: 1) the preservation concepts are explicitly referred to, which increases the understanding of the design, 2) the same preservation concepts can be reused across multiple descriptions, 3) the different elements of the description can be consistently described and reused under different perspectives, and 4) parts of the formal description can lead to software implementation using a model-driven architecture approach.

The DoDAF metamodel is implemented as an extension of the Unified Modeling Language (UML)[19] in the Unified Profile for DoDAF/MODAF (UPDM) [20]. UPDM makes it possible to develop architectural descriptions with generic UML modeling tools. DoDAF concepts are implemented as stereotypes, and views are implemented as UML diagrams.

3 Approach to the architectural description of archives

The approach presented in this paper relies on enterprise architecture to describe archival systems, to formally describe the interactions involving the archival system, and the functions performed by this system. Although the approach does not focus on a particular content type, it can be used to represent, in a coherent fashion, the complex interactions and information models involved in product data preservation. The intent is to provide a high-level description of the archival system, which can then drive the actual software implementation of the whole system.

This approach, depicted in Figure 2, is split into two parts. The first part extends the metamodel defined in DoDAF to incorporate a new archival vocabulary derived from the OAIS RM. The expanded terminology can then be used to represent archival system elements in architectural descriptions. The second part selects DoDAF views to represent specific aspects of the preservation solution. These views can serve as ACTDR evidence to demonstrate the ability of

the archival system to preserve information. The approach, which is referred to as Reference Architecture for Archival System, can serve to guide and constrain architectural descriptions of archival systems. Note that a complete archival system description would go beyond what is in the scope of our approach, so other views may be used to address other aspects.

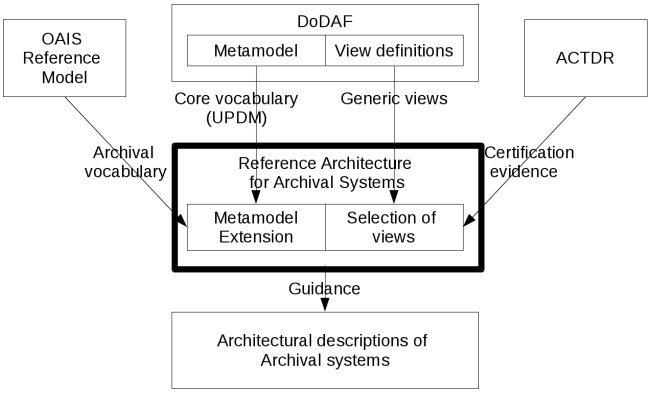

Fig. 2. Presentation of the approach

3.1 Representation of preservation concepts

DoDAF provides a generic, formal vocabulary to describe architectures, and the OAIS RM provides an informal, specific vocabulary to describe archival systems. Our approach combines both to allow a formal description of archival systems using enterprise architecture, and by incorporating the preservation terminology. Only the terminology related to the archival system and the interactions between this system and the environment are considered. Preservation functions such as customer monitoring or technology monitoring are present in the OAIS RM, but they are not in the scope of our approach. While these functions can be represented using DoDAF, we focus on the interactions between the archival system and the other business functions.

The OAIS RM describes archives in a conceptual and technology-independent manner, while DoDAF describes concrete implementations within an organization. So, more preservation concepts can be inferred by determining what DoDAF concepts would be used in an archival system description.

The approach uses UPDM, an implementation of DoDAF as a UML profile, which allows using the DoDAF terminology in UML tools. We will provide a conceptual description of the approach, as opposed to a detailed implementation.

The core terminology of DoDAF used in our approach is presented in Figure 3. The Figure shows how the different enterprise architecture concepts relate to each other. A *System* provides *Services*, and performs *Functions*. A *Node*

performs *Activities*. Both *Functions* and *Activities* involve *Information*. A *Standard* can apply to *Information*, and *Constraints* can apply to *Information* or *Activities*.

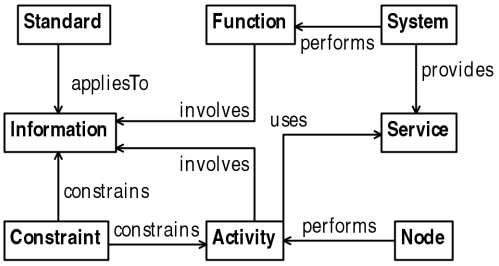

Fig. 3. DoDAF concepts used in the approach

3.2 Adapting the DoDAF concepts for preservation

The following paragraphs present some of the preservation concepts considered in our approach. The objective is to provide a way to formally describe the preserved content, the information packages, the representation information, the preservation description information, the operational nodes, the activities, the system functions, the constraints, and the standards. The preservation concepts are written in italics.

Content and information packages Content and information packages are OAIS concepts relating to information that is exchanged during preservation-related activities. Content is the information that is meant to be preserved. Information packages encapsulate content information as well as additional information required to ensure a long-term preservation and accessibility. The SIP, AIP, and DIP represent the information packages respectively as they are received, preserved, and disseminated. In UPDM, the concept *ExchangeElement* represents a resource exchanged during an activity, so both content and information packages are defined as the following specializations of *ExchangeElement*: *SubmissionInformationPackage*, *ArchivalInformationPackage*, *DisseminationInformationPackage*, and *Content*.

In the context of product data preservation, the *SubmissionInformationPackage*, the *ArchivalInformationPackage*, and the *DisseminationInformationPackage* represent the container of product data during ingest, retention, and dissemination activities respectively. A *Content* may represent the actual product data, or a piece of information that corresponds to the target of the preservation.

Representation information and presentation description information Within information packages, the OAIS RM defines Preservation Description Information (PDI) and representation information that can be attached to the content. Representation information represents information that gives more meaning to the data: it could be everything that allows computers or humans to interpret the data, such as format specifications, software, or dictionaries. PDI is further detailed in four categories. Reference information identifies the content. Provenance information describes the content history. Fixity information represents the information used to check that the content is not altered. Finally, Context information provides the relationships between a content and the other various contents.

In the context of product data preservation, some of the preservation description information can be extracted from Product Data Management (PDM) or Product Lifecycle Management (PLM) systems, such as identifiers, creators, or relationships among product data.

Nodes Producers and consumers can also be seen as *nodes* instead of physical persons. A *node* is a logical abstraction, meaning that it may correspond to people or systems. The following specializations of nodes are added *Producer*, *Consumer*, *Preserver*, and *Archive*.

In the context of product data preservation, *Producer* may correspond to the data source from which the product data originate (e.g., PDM systems), *Archive* abstracts the physical realization of the archival system, and *Consumer* may correspond to where the product data is used over time. *Preserver* can represent the persons in charge of the preservation of product data.

Activities Three kinds of activities can be identified: the interaction between the archive and the producers, consumers, and management constitute respectively ingest, access, and management activities. In addition, the activities that are within the OAIS are also defined, in particular the preservation activities that include update and disposal of the preserved content. All of these activities are defined as specialization of *OperationalActivities* in UPDM: *IngestActivity*, *PreservationActivity*, and *AccessActivity*.

In the context of product data preservation, *IngestActivity* may represent the activities of taking product data from their original place, preparing a *SubmissionInformationPackage*, and sending it to the archive. *PreservationActivity* may represent the activities undertaken for preserving the product data over time by accessing the archival system. *AccessActivity* may represent the activities that request product data from the archive.

Services Services constitute another concept that is important in the implementation of archival systems. Nowadays many software development approaches rely on a Service-Oriented Architecture. UPDM supports this approach by defining the notion of *service*. *IngestServices* are the services exposed to the producers for the ingest. *ManagementServices* are the services exposed to the preservers

to make sure the content stays interpretable. *AccessServices* are the services exposed to the consumers for accessing the preserved content.

System functions The OAIS RM defines various functions that an OAIS performs. Two kinds of functions can actually be differentiated: those intended to be performed by humans, and those intended to be performed by computers. UPDM makes the distinction between these two types, and calls them respectively *OperationalActivities* and *Functions*. The *Functions* that are likely to be implemented by systems are the following: *IngestFunction* ingests the content, *ManagementFunction* manages the preserved content, and *AccessFunctions* makes the preserved content accessible.

4 Conclusion

This paper discussed an approach to formally describe the information and activities related to archival systems. This approach relies on the DoDAF enterprise architecture framework, and it uses a preservation terminology inspired by the Reference Model for an Open Archival Information System to describe the main elements of the archival system. Using this approach, the preservation concepts defined in the OAIS RM can be referred to within archival system designs. DoDAF also provides a way to consistently define and interrelate the different elements that constitute an archival system. This approach can lead to the definition of a comprehensive reference architecture for archival systems. A caveat is that to maximize the approach's usefulness, the entire enterprise should be described using DoDAF.

This approach is generic enough to be used in many different preservation cases, including product data preservation. The long-term access of product data is essential for product lifecycle management, especially in the case of products with a long life. This approach can be demonstrated by describing a product data ingest, to show the activities and information involved in the ingest. The ingest activity can be formally described according to the OAIS RM terminology, showing for example the transfer from PDM systems to the archive. From the information perspective, the content and the metadata can also be formally defined.

Disclaimer No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial software are identified in this report to facilitate better understanding. Such identification does not imply recommendations or endorsement by NIST nor does it imply the software identified are necessarily the best available for the purpose.

References

[1] Neil Beagrie. Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, 1(1):3–16, autumn 2006.

[2] T Kuipers and JVD Hoeven. Insight into digital preservation of research output in europe – survey report, 2009.
[3] Blue Ribbon Task Force. Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information, 2010.
[4] Manjula Patel, Alexander Ball, and Lian Ding. Strategies for the curation of cad engineering models. *International Journal of Digital Curation*, 4(1): 84–97, 2009. ISSN 1746-8256.
[5] William C. Regli, Michael Grauer, and Joseph B. Kopena. A framework for preservable geometry-centric artifacts. In *2009 SIAM/ACM Joint Conference on Geometric and Physical Modeling*, pages 67–78, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-711-0. doi: http://doi.acm.org/10.1145/1629255.1629265.
[6] International Organization for Standardization. ISO 14721:2012, Space data and information transfer systems – Open archival information system (OAIS) – Reference model, 2012.
[7] ASD-STAN. LOTAR - LOng Term Archiving and Retrieval of digital technical product documentation such as 3D, CAD and PDM data.
[8] Verband der Automobilindustrie e. V. VDA 4958: Long-Term Archiving of digital Product Data, which are not based on technical drawings. Technical report, Verband der Automobilindustrie e. V., 2005.
[9] Consultative Committee for Space Data Systems. Audit and Certification of Trustworthy Digital Repositories, October 2009.
[10] S.A. Bernard. *An introduction to enterprise architecture*. AuthorHouse, 2005.
[11] The Open Group, editor. *TOGAF Version 9.1, Enterprise Edition*. Van Haren Publishing, 2011.
[12] Ministry of Defence Architecture Framework (MODAF) 1.2.
[13] Department of Defense Architecture Framework (DoDAF) 2.0.
[14] IFIP-IFAC Task Force. IFIP-IFAC task force on architectures for integrating manufacturing activities and enterprises. In *IFIP newsletter/IFAC newsletter*. 1993.
[15] International Organization for Standardization. ISO/IS 15704:2000, industrial automation systems requirements for enterprise reference architectures and methodologies, 2000.
[16] T.J. Williams. The Purdue enterprise reference architecture. *Computers in industry*, 24(2-3):141–158, 1994.
[17] A CIMOSA presentation of an integrated product design review framework. *Int. J. Computer Integrated Manufacturing*, 84(4):260–278, 2005.
[18] C. Becker, G. Antunes, J. Barateiro, R. Vieira, and J. Borbinha. Modeling digital preservation capabilities in enterprise architecture. In *In 12th Annual International Conference on Digital Government Research (dg. o 2011), June 12-15, College Park, MD, USA*, 2011.
[19] Unified Modeling Language (UML) 2.0.
[20] Unified Profile for the Department of Defense Architecture Framework (DoDAF) and the Ministry of Defence Architecture Framework (MODAF).