DOCUMENT 2: ON THE FUNCTIONING AND CHARACTERISTICS OF THE LEGATUM REPOSITORY - SONUS ET IMAGO

a)   HOW DOES THE LEGATUM REPOSITORY WORK?

The Legatum Repository aims to establish itself as a federated environment, and already brings together a number of partner and collaborating public institutions. The Legatum Repository is characterized by two interconnected environments, one for *access* and one for *digital preservation*. For remote access, it's used the software *Access to Memory (AtoM)*, and for digital preservation, the *Archivematica* software, both of which allow us to comply to Brazilian legislation recommendations, regarding access and preservation of documents produced by public institutions. In the Legatum Repository they will both be interconnected, allowing us to perform digital preservation and generate reliable access versions. These software are free and open source, and are recognized by the International Council of Archives (ICA) and the National Council of Archives of Brazil (CONARQ). They are currently adopted, for example, by UNESCO, among many other important institutions in many countries. AtoM and Archivematica are based on international archival standards for digital preservation and archival remote access.

Thus, in the AtoM environment of the Legatum Repository, the metadata schema adopts the General International Standard Archival Description (ISAD-G), also seeking to comply with the Brazilian Standard for Archival Description (NOBRADE), the International Standard for Describing Institutions with Archival Holdings (ISDIAH), the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR-CPF) and the International Standard for Describing Functions (ISDF).

Archivematica, in turn, was developed based on ISO 14.721:2012 - Space data and information transfer systems - Open archival information system (OAIS) - Reference model. Due to the importance that this standard is acquiring all over the world and to the fact that the Archivematica software adopts this model, we will go into a little more detail here.

The OAIS Reference Model (translated in ABNT NBR 15.472: 2007[1] as *Sistema Aberto de Arquivamento de Informação* - SAAI).

> provides a systematic framework for understanding and implementing the archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives. It describes roles, processes and methods for long-term preservation. Developed by the Consultative Committee for Space Data Systems (CCSDS) OAIS

---

[1] ABNT 15.472: 2007 is based on ISO 14.721: 2003, (practically a Portuguese version in its entirety), and so, it has currently a gap of almost 15 years in relation to the new version of ISO, since in January 9, 2018, it has been revised and validated.

was first published in 1999 and has had an influence upon many digital preservation developments since the early 2000s.[2]

According to the standard NBR 15.472:2007 (p.vi):

> The Open Archival Information System (OAIS) is a file, understood in this Standard, as an organization of people and systems, which has accepted the responsibility of preserving information and making it available to a target community. An OAIS file stands out from other uses of the term 'file' because it fulfills a set of responsibilities, as the ones defined in this Standard. The term 'open' in OAIS means that this recommendation, as well as future recommendations and related standards, are to be developed in open forums, but it does not mean that access to the file is unrestricted.[3]

Also according to the previous NBR standard, the OAIS model:

- provides a framework for a broad understanding and increased awareness of archival concepts needed for Long Term digital information preservation and access;
- provides the concepts needed by non-archival organizations to be effective participants in the preservation process;
- provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future Archives;
- provides a framework for describing and comparing different Long Term Preservation strategies and techniques;
- provides a basis for comparing provides a basis for comparing the data models of digital information preserved by Archives and for discussing how data models and the underlying information may change over time;
- provides a framework that may be expanded by other efforts efforts to cover Long Term Preservation of information that is NOT in digital form (e.g., physical media and physical samples) …[4]

We are in Phase 3 of 3 of the survey, which is expected to be completed in February 2019. The active and continuous participation of partner and collaborating institutions, even after the research is completed, is essential for the evolution of the Legatum Repository and, in consequence, for the digital preservation of public audiovisual collections, which have grown exponentially in recent years. UFBA, in partnership with the Brazilian Institute of Information on Science and Technology (IBICT), can lead this innovation.

---

[2] Cf. Digital Preservation Coalition. Available at: https://dpconline.org/handbook/institutional-strategies/standards-and-best-practice. Access in: 23jun2018

[3] Translation based on the original ISO 14.721.

[4] Translation based on the original ISO 14.721:2012. In the original, published in English, there are other two aspects: a) expands consensus on the elements and processes for Long Term digital information preservation and access, and promotes a larger market which vendors can support; and b) guides the identification and production of OAIS-related standards.

b) **WHAT CONTENT DOES THE LEGATUM REPOSITORY MAKE AVAILABLE?**

Repositories that adopt the AtoM and Archivematica software support any type of digital document in their environments. However, the Legatum Repository is dedicated to the complex digital audiovisual archives (and also iconographic, textual, photographic, as long as they relate to the accepted audiovisual archive). Thus, digital born versions or resulting from processes of digital conversion of items from audiovisual collections of an institutional, scientific, academic and artistic nature, can be made available on this platform.

c) **WHAT IS THE INFRASTRUCTURE AND PLATFORM OF THE LEGATUM REPOSITORY?**

The Legatum Repository platform for the access function is hosted in a virtual machine provided by the Superintendence of Information Technology of the Federal University of Bahia (STI-UFBA). The operating system on the virtual machine is the Debian Linux distribution, stable version. The application is based on the free software AtoM, currently in version 2.4.0. The programming language is PHP.

The Legatum Repository module for the Digital Preservation function is hosted on a dedicated server computer, provided by Dell and acquired with scholarship resources in productivity from the National Council for Scientific and Technological Development (CNPq). The operating system is the Linux distribution Ubuntu 14.04, Long Term Support (LTS). The application is based on the free software Archivematica, whose programming language is Python.

The master files (original born digital or digital representatives intended for digital preservation), at the present time, will not go online on the server where Archivematica digital preservation software is installed. They will be recorded on Linear Tape Open (LTO), in three copies, so that 01 copy stays with the partner/collaborator institution, 01 copy stays with an entity indicated by the partner/collaborator institution and 01 copy stays with UFBA. However, we are studying the advantages of cloud computing infrastructures, since they have already been adopted by UFBA, internally. We are thinking, for a cooperative future, about the possibility of expanding it in a collaborative way with other prominent public institutions. By adopting "the cloud", the work with infrastructure is reduced. To record and maintain LTO tapes, is also dealing with infrastructure, at this point of the research, but – with the possibility of hiring external cloud infrastructures (public, institutional, cooperatives, distributed) – issues associated with hardware acquisitions, media and software upgrades, would no longer be a difficulty for partner institutions and for UFBA. By way of illustration, it is known to exist, today, in 2018, options with cost of storage at USD 0.01 per GB/month (or $ 1 per Terabyte), but a more dense analysis is needed and other aspects of the public

sphere have to be considered. However, the cloud storage option cannot be overlooked, since all infrastructure work is passed on/outsourced to the company providing the service. In this case, instead of maintaining an infrastructure within the institution, we would have only configuration, upload and download, when necessary. On the other hand, the cloud solution involves other issues, not just costs (especially for high-resolution digital video files that require considerable storage space), such as being dependent on a company, which may close without previous notice. Being Legatum a public initiative, that's something to take into account. But the fact is that there is a new field to be considered in this line, not just about storage, but a host of other features, such as artificial intelligence (AI), which allows you to automatically analyze videos and index them (what is shown in each moment of the video). We have a new generation of AI-based access systems that are very close to reaching the market, and we have no doubt that AI will also help in digital preservation, very soon. Another aspect of our current line of thought is based on the premise that you cannot save everything: there may be need for more strict criteria for evaluation and selection, such as considering only the final versions of the works and not all the material created during the production. The difficult thing would be to define criteria to categorize audiovisual archives separating, for example, masterpieces from the most common ones, or those that ideally represent a time or a technique, that have been chosen for preservation. Regarding the definition of these criteria, this has been an issue that has historically been a complex subject for audiovisual archives. Therefore, such a solution should be established by the members of the Group of Studies on Culture, Representation and Digital Information (CRIDI), its formal partners and UFBA, through a document that would reflect the consensus, amidst the difficulties that characterize how far one can go, always considering current limitations.