DEPENDENCE RESCISSION AND SEMANTIC INTERPRETATION: KEY STRATEGIES TO LONG-TERM PRESERVATION OF DIGITAL ARCHIVES

Yi Qian
Renmin University of China
China
Qianyi1973@126.com
ORCID 0000-0003-0609-0791

Linqing Ma
Renmin University of China
China
Malinqing2010@126.com
ORCID 0000-0001-9455-9750

Abstract - There are three states of archival objects and their corresponding management measures. Clarifying the continuity or discreteness of signals and semantics of three states is very important in preservation. For objects at digital state, low-order logic dependencies with software and hardware dependencies as the core need to be removed. And for objects at data state, high-order logic dependencies with semantic associations as the core need to be removed.

Keywords - analog state,digital state, data state, dependency removal, semantic interpretation

Conference Topics – Sub-theme 1: Exploring the New Horizons.

I. NEW CHANGES: THREE STATES OF ARCHIVAL OBJECTS AND THEIR CORRESPONDING MANAGEMENT MEASURES

The evolution of technology environments have significantly affected archival objects. There are three generations of technical environments, i.e., the traditional analog technology environment, the digital technology environment based on digital signals, and the data technology environment with the core of data-driving. Accordingly, the archival objects created in the three generations of technology environments can be referred to as analog state, digital state and data state[1].

Archival objects in the three states have obvious distinctions, mainly in the nature of the signals and semantic continuity of the content. Archival objects in analog state are recorded and preserved in analog signals (continuous physical signals) and in the form of documents, such as various traditional paper documents, photos and so on. Archival objects in digital state are recorded and preserved in discrete digital signals and take the computer files as containers, such as general digital records and digital photos. Archival objects in data state are recorded and preserved in digital signals and the granularity of recorded information is at the level of data, such as relational database, GIS data, and 3D data. Archival objects in data state are various data and data sets which are generated by data-driven systems and oriented for machine-processing.

The archival objects are undergoing the change from analog state and digital state to data state. Accordingly, the management measures for archival objects should also change, which include the way of thinking, management environments, technological tools, and preservation strategies.

TABLE I
Comparison of Features of Management Measures Based on Three States Theory[2]

	Analog state	Digital state	Data state
Nature of signals	Analog (continuous)	Digital (discrete)	Digital (discrete)
Semantic features	Continuous semantics	Continuous semantics	Discrete semantics
Managed objects	Carrier-centric	Content-centric	Data-centric
Dependencies	Unity of logical structure and physical structure	Separation of information (logical structure) from carrier (physical structure)	Logical structures and models with complex associations
structure of Objects	Carrier + handwriting + combination	Content, structure, context	Related elements such as rules, models, and semantics
Original form	Paper + handwriting	Solidified content within	Data body + parameters

17th International Conference on Digital Preservation
iPRES 2020, Beijing,China.

DOI: 10.1145/nnnnnnn.nnnnnnn

		a record	snapshot
Focus of Management	Orderly environmental management based on the carrier	Management system based on four characteristics of records	rule, model, and ontology management based on data
Technical measures	Elements protection	System management	Smart Service
Completeness of regulation	Almost complete	Under construction	Blank

II. CHALLENGES TO LONG-TERM PRESERVATION BROUGHT BY THE THREE STATES

For archival objects at the analog state, since the analog signal is continuous and can be received and read by humans directly, and the format is also human-readable and understandable, we say that the characteristics of archival objects at the analog state are signal continuity and semantic continuity. For archival objects at the digital state, since documents are taken as containers for semantic encapsulation, the definitions, elements, and overall construction of documents are designed based on semantic understanding requirements of humans, they still have semantic continuity, although the signals are discrete.

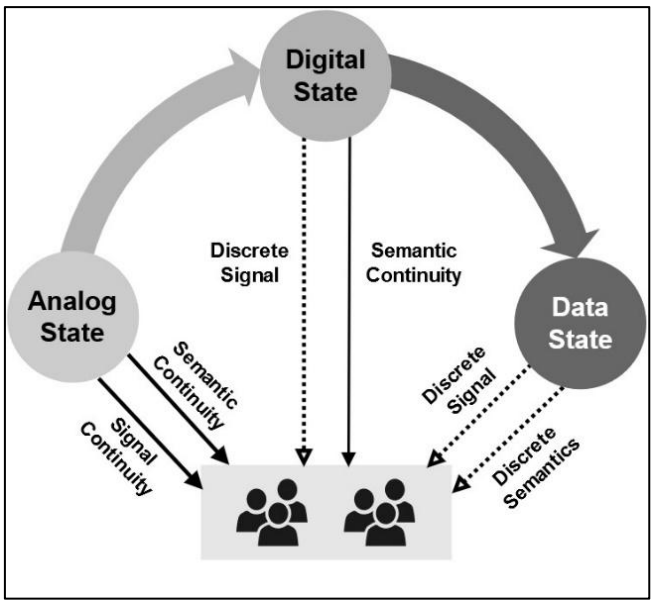

Figure 1 Features of Archival Objects of Three States[3]

However, archival objects at the data state are in the different situation, and having the characteristics of discrete signals and discrete semantics. To deal with the challenges of discrete signals, software and hardware dependencies need to be rescinded. And to deal with the challenges of semantic discreteness, content (data) interpretation tools are needed. In practice, we use human-readable and machine-readable to describe these measures. In terms of the ability to read discrete/continuous signals, humans and machines can be regarded as two extremes in principle. Human can only read the materials of signal continuity and semantic continuity. Whereas, machine can read materials in discrete signals and semantics.

The essential feature of data state is that both signals and semantics are discrete. In particular, the discreteness of semantics complicates the preservation issue. The core of data state preservation is to maintain the original business expressed by the semantics and rules of discrete data. In digital state, content-based unstructured files generally do not have the problem of semantic discreteness, because the documents with semantic encapsulation formats can provide support for semantic understanding to a considerable extent. For the preservation of semantic discreteness, it is necessary to adopt corresponding semantic integration methods to improve semantic capabilities like interpretation, decoding, integration, and presentation, such as view tools in databases and data exchange standards in the preservation of complex models. But these tools themselves need a high level of usage to achieve semantic continuity.

III. TYPICAL STRATEGIES TO ARCHIVAL OBJECTS PRESERVATION AT THE THREE STATES

Due to the differences of archival objects at different states, there are differences in dependency levels. For each state, the dependency level that the preservation problem needs to be resolved is different, that is, the Benchmarks of dependency of the three states are different. The so-called Benchmarks of dependency refers to the lowest level of dependency that needs to be solved to achieve the intelligibility goal in different states. For example, the Benchmark of dependency at analog state is carrier dependency which means that we can understand object by storing object's carrier, and the Benchmark of dependency at digital state is grammatical dependency which means that we need to use machines (software and hardware facilities) to translate code into information we can read. The above two dependencies can be regarded as low-level logical dependency. However, at least some semantic dependency need to be resolved at data state because of signal discretization and semantic discretization. This dependency can be regarded as high-level logical dependency. Therefore, the preservation levels of the three states as a whole are improved along the progressive logic of "visible-readable-understandable". Due to the hierarchical superposition of dependency, it is necessary to establish corresponding Benchmarks of dependency at different states. The preservation levels of archival

objects are determined by the benchmarks of dependency and the designed communities.

A. Environmental Control: Analog state preservation mainly through physical, and bio-chemical methods

Archival objects at analog signals need to resolve carrier dependency. The main measures are to deal with the physical preservation of the carrier through physical, chemical, biological, to ensure that the carrier itself is visible. Its typical practice is so called "eight prevention measures" for archival repository in archival field, which are related with light, heat, fire, moisture, dust, rats, insects and so on.

B. Dependency releasing: Digital state preservation focusing on releasing software and hardware dependency

Archival objects at digital state need to solve the low-level logic dependency of software and hardware as the core, and deal with the preservation problem caused by the discrete signals. Focusing on software and hardware dependency issues, to solve the problems of electronic records reading relying on hardware devices, character recognition relying on operating systems, and content browsing relying on related software to ensure that electronic records are readable, specific measures are needed, such as format management, metadata management, migration, and construction of the TDR(Trusted Digital Repository) systems.

C. Semantic interpretation: Data state preservation focusing on semantics

Archival objects at data state need to solve the high-level logical dependency with semantic association as the core, and the purpose is to deal with the preservation problem caused by signal discretization and semantic discretization. It can be considered as preserving algorithms, rules, constraints, models which can help to understand semantics and their expressions.

Typical measures include using views to archive and preserve dynamic data. The core role of this approach is to integrate the discrete semantics of dynamic data in a view to form a continuous semantic state that humans can understand, such as reintegrating the original data scattered in the tables of the database into views through data relationships. The standard ISO 16175-3(Information and documentation — Principles and functional requirements for records in electronic office environments -Part3: Guidelines and functional requirements for records in business systems) illustrates more on it. In the examples provided by this standard, a digital record is made up of related data elements from different data tables. To fully understand this standard, in addition to the basic data, it is required to provide the necessary structure and context from the relational databases, standard data modeling and normalization techniques, such as primary keys, foreign keys, stored procedures, various constraints and so on, to ensure the completeness of the semantics and traceability of digital records. Currently, there is no clear approach to archive electronic records in the form of databases. In the field of long-term preservation, semi-structured methods (normally XML) are usually used to preserve structured databases. The database archiving format standard proposed by the Swiss Federal Archives' SIARD (Software Independent Archiving of Relational Databases)project is representative[4].

REFERENCES

[] Yi. Qian, "Evolution of archival objects management measures in technology change," *Archives Science Bulletin*, no. 2, pp.10-14, 2018.

[2]Yi. Qian, "From 'digitalization' to 'datamation'-new understanding of several issues in record management in the new technology environment," *Archives Science Bulletin*, no. 5, pp.42-45, 2018.

[3]Yi. Qian, "From 'digitalization' to 'datamation'-new understanding of several issues in record management in the new technology environment," Archives Science Bulletin, no. 5, pp.42-45, 2018.

[4]Swiss Federal Archives. SIARD (Software Independent Archiving of Relational Databases) Version 1.0[EB/OL]. [2020-01-15]. https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml.