Digital Archives: Extending the 5S model through NESTOR

Nicola Ferro and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{ferro, silvello}@dei.unipd.it

Abstract. Archives are an extremely valuable part of our cultural heritage. Although their importance, the models and technologies that have been developed over the past two decades in the *Digital Library (DL)* field have not been specifically tailored on archives and this is especially true when it comes to formal and foundational frameworks, as the *Streams, Structures, Spaces, Scenarios, Societies (5S)* model is. Therefore, we propose an innovative formal model, called *NEsted SeTs for Object hieRarchies (NESTOR)*, for archives, using it to extend the 5S model in order to take into account the specific features of the archives and to tailor the notion of digital library accordingly.

1 Motivation

Over the past two decades, *Digital Libraries (DLs)* have been steadily evolving and have been shaping the way in which people and institutions access to and interact with our cultural heritage, study, and learn. Nowadays, their reach goes far beyond what has been the realm of traditional libraries and encompasses also other kinds of cultural heritage institutions, such as archives and museums. In particular, this work focuses on archives; an archive is not simply constituted by a series of objects that have been accumulated and filed with the passing of time – as it usually happens with libraries that collect, for example, individual published books, journals, and serials. Instead, it represents the trace of the activities of a physical or juridical person in the course of their business which is preserved because of their continued value.

DLs benefit from the existence of sophisticated formal models, such as the *Streams, Structures, Spaces, Scenarios, Societies (5S)* model [4], which allow us to formally describe them and to prove their properties and features. Notwithstanding the importance of the archives, so far, there has been no attempt to develop a dedicated formal model, built around their peculiar constituents, such as the notion of *archival bond*. We can neither exploit the 5S model as it is for archives because, as we will discuss later on, it needs some kind of extension and tailoring.

We think that the archival domain deserves a formal theory as well and that this theory has to be reconciled with the more general theories for digital libraries in order to disclose to archives the full breadth of methodologies and technologies which have been developed over the last two decades in the DL field. To this purpose we proposed a formal model for archives, built around the notion of *archival bond* and *hierarchy*: the *NEsted SeTs for Object hieRarchies (NESTOR)* model [1]. Furthermore, we exploit NESTOR to formally extend the 5S model in order to be capable of defining a *digital archive* as a specific case of digital library able to take into consideration the peculiar features of the archives.

The paper is organized as follows: in Section 2 we provide some background on archives and the 5S formal model. In Section 3 we present the basics of the NESTOR model and in Section 4 we introduce our extension to the 5S model via NESTOR. Finally, in Section 5 we draw some final remarks.

2 Related Work

2.1 Digital Archives

In an archive the context and the relationships between the documents are preserved thanks to the hierarchical organization of the documents inside the archive. Indeed, an archive is divided by fonds and then by sub-fonds and then by series and so on; at every level we can find documents belonging to a particular division of the archive or documents describing the nature of the considered level of the archive. The union of all these documents, the relationships and the context information permits the full informational power of the archival documents to be maintained. The archival documents are analyzed, organized, and recorded by means of the *archival descriptions* that have to reflect the peculiarities of the archive. In the digital environment archival descriptions are encoded by the use of metadata; these need to be able to express and maintain the structure of the descriptions and their relationships [3].

The standard format of metadata for representing the hierarchical structure of the archive is the *Encoded Archival Description (EAD)*[1], which reflects the archival structure and holds relations between entities in an archive. On the other hand, an archive is described by means of a unique EAD file and this may be problematic when we need to access and exchange archival metadata with a variable granularity [2].

2.2 The 5S Model

The *Streams, Structures, Spaces, Scenarios, Societies (5S)* [4] is a formal model and draws upon the broad DL literature in order to have a comprehensive base of support. It has been developed largely bottom up, starting with key definitions and with elucidation of the DL concepts from a minimalist approach. It is built around five main concepts:

- *streams* are sequences of elements of an arbitrary type, e.g. bits, character, images, and so on;

[1] `http://www.loc.gov/ead/`

- *structures* specify the way in which parts of a whole are arranged or organized, e.g. hypertexts, taxonomies, and so on;
- *spaces* are sets of objects together with operations on those objects that obey certain constraints, e.g. vector spaces, probabilistic spaces, and so on;
- *scenarios* are sequences of related transition events, for instance, a story that describes possible ways to use a system to accomplish some functions that a user desires;
- *societies* are sets of entities and relationships between them, e.g. humans, hardware and software components, and so on.

Starting from these five main concepts, it provides a definition for a minimal DL which is constituted by: (i) a repository of digital objects; (ii) a set of metadata catalogs containing metadata specifications for those digital objects; (iii) a set of services containing at least services for indexing, searching, and browsing; and, (iv) a society.

While these broad concepts can be in common also with archives, when you look at the specific way in which they are formally defined, you realize that the definitions cannot be straightforwardly applied to the archives case without at least some extension. We will discuss this in further details, presenting an extension of 5S via NESTOR in Section 4.

3 The Basics of the NESTOR Formal Model

We define both *Nested Sets Model (NS-M)* and *Inverse Nested Sets Model (INS-M)* in terms of the set theory as a collection of subsets where specific conditions must hold.

Definition 1 *Let A be a set and let $\mathcal{C}$ be a collection of subsets of A. Then $\mathcal{C}$ is a* ***Nested Sets Collection*** *(NS-C) if:*

$$A \in \mathcal{C}, \tag{3.1}$$

$$\forall H, K \in \mathcal{C} \mid H \cap K \neq \emptyset \Rightarrow H \subseteq K \vee K \subseteq H. \tag{3.2}$$

Therefore, we define a NS-C as a collection of subsets where two conditions must hold. The first condition (3.1) states that set A which contains all the subsets of the collection must belong to the NS-C itself. The second condition states the intersection of every couple of sets in the NS-C is not the empty-set only if one set is a proper subset of the other one.

Now we can introduce the Inverse Nested Sets Collection (INS-C) which defines the INS-M:

Definition 2 *Let A be a set and let $\mathcal{C}$ be a collection. Then, $\mathcal{C}$ is an* ***Inverse Nested Sets Collection*** *(INS-C) if:*

$$\exists! B \in \mathcal{C} \mid \forall K \in \mathcal{C}, B \subseteq K, \tag{3.3}$$

$$\forall H, K, L \in \mathcal{C} \mid H \subseteq K, L \neq K \Rightarrow (L \cap K = H \cap L) \vee (H \subseteq L) \vee (L \subseteq H). \tag{3.4}$$

We define an INS-C as a collection of subsets where two conditions must hold. The first condition (3.3) states that $\mathcal{C}$ must contain the *bottom set* B, which is the common subset of all the sets in $\mathcal{C}$. The second condition (3.4) states that if we consider three sets K, H, and L such that H is a subset of K and K is not equal to L, then the intersection between L and K is not the same as the intersection between H and L or H is not a subset of L and vice versa.

4 Extending the 5S Model via NESTOR

The notion of *descriptive metadata specification*[2] (definition 14 [4, p. 293]) is suitable either to represent, for each archival division, a descriptive metadata – e.g. a metadata describing a serie, a sub-fonds, or an archival unit – or to represent the archive as a whole, as it happens in the case of EAD.

When it comes to the definition of *metadata catalog* (definition 18 [4, p. 295]), there is no means to impose a structure over the descriptive metadata in the catalog. Therefore, if you use separate *descriptive metadata specifications* for each archival division, as in the former case, this would prevent the possibility of expressing the relationships between these archival divisions, i.e. you would loose the possibility of retaining the archival bond.

Moreover, in a *metadata catalog*, there is no means to associate (sub-)parts of the *descriptive metadata specifications* to the *digital objects* (definition 16 [4, p. 294]) that they describe, but you can only associate a whole descriptive metadata to a whole digital object.

Therefore, if you represent an archive as a whole with a single *descriptive metadata specification*, as in the latter case, it would not be possible to associate (sub-)parts of that descriptive metadata to the different digital objects corresponding to the various archival divisions.

Our extension to the 5S model is thus organized as follows:

- using the notion of *structure* (definition 2 [4, p. 288]), we introduce the notion of **NESTOR structure**, as a structure that complies with the constraints of NS-M or INS-M;
- using the notion of *metadata catalog*, we introduce the notion of **NESTOR metadata catalog**, as a metadata catalog that exploits a NESTOR structure to retain the archival bonds;
- using the notion of *digital library* (definition 24 [4, p. 299]), we introduce the notion of **digital archive**, as a digital library where at least one of the *metadata catalogs* is a NESTOR metadata catalog.

Definition 3 *Let $\mathcal{C}$ be a Nested Set Collection (NS-C) on a set A. A* ***NS-M structure***(A) *is a structure* $(NS\text{-}G, L, \mathcal{F})$*, where L is a set of label values, $\mathcal{F}$ is a labeling function, and* $NS\text{-}G = (V, E)$ *is a directed graph where* $\forall v_j \in V, \exists!\ J \in \mathcal{C} \wedge \forall e_{j,k} \in E, \exists!\ J, K \in \mathcal{C} \mid K \subseteq J$.

[2] In this section, we use italics for highlighting definitions taken from the 5S model.

Definition 4 *Let $\mathcal{C}$ be an Inverse Nested Set Collection (INS-C) on a set A. A* ***INS-M structure****(A) is a structure $(INS\text{-}G, L, \mathcal{F})$, where L is a set of label values, $\mathcal{F}$ is a labeling function, and $INS\text{-}G = (V, E)$ is a directed graph where $\forall v_j \in V, \exists! J \in \mathcal{C} \wedge \forall e_{j,k} \in E, \exists! J, K \in \mathcal{C} \mid J \subseteq K$.*

Definition 3 applies definition 1, ensuring that the resulting structure complies with the NS-M. Note that the set of label values L and the labeling function $\mathcal{F}$ are not strictly needed for the NS-M, but they can be useful in the context of the 5S and this feature, in turn, may extend the NS-M with semantic possibilities. Similarly, definition 4 applies definition 2.

Definition 5 *Given a set A, a* ***NESTOR structure****(A) is either a NS-M structure(A) or a INS-M structure(A).*

The definition of *metadata catalog* in the 5S model can be expressed as follows. Let H be a set of handles to *digital objects* and M a set of *descriptive metadata specifications*, then a *metadata catalog* is a function $DM : H \times 2^M$.

Definition 6 *Let H be a set of handles to digital objects and M a set of descriptive metadata specifications, a metadata catalog DM is a* ***NESTOR metadata catalog*** *if:*

$$\forall h_i \in H \mid \exists M_i \in 2^M \ \wedge \ DM(h_i) = M_i \Rightarrow |M_i| = 1 \tag{4.1}$$

$$\exists \text{ NESTOR structure}(M) \tag{4.2}$$

Condition 4.1 imposes that, if exists, there is only one *descriptive metadata specification* for a given *digital object* because, in the archival practice, every single metadata describes a unique archival division, being it a level in the archive or a digital object [5]. Condition 4.2 ensures that the relationships among the different archival divisions are compliant with the *descriptive metadata specifications* in M.

Definition 7 *A* ***digital archive*** $(\mathcal{R}, DM, \mathrm{Serv}, Soc)$ *is a digital library where*

- *$\mathcal{R}$ is a repository;*
- *at least one of the metadata catalogs in the set of metadata catalogs DM is a NESTOR metadata catalog;*
- *Serv is a set of services containing at least services for indexing, searching, and browsing;*
- *Soc is a society.*

Definition 7 extends the definition of *digital library* in the 5S model requiring that at least one of the *metadata catalog* is a NESTOR one, i.e. there exists at least on *metadata catalog* capable of retaining the archival bonds.

5 Final Remarks

The definition of digital archive we gave in this paper has a couple of consequences. Firstly, more NESTOR metadata catalogs can be present in the same digital archive, thus giving the possibility of expressing different archival descriptions over the same set of *digital objects*. This extends the current practice in which a system for managing an archive is usually capable of managing only one description of the archive, thus giving only one point-of-view on the held material. Secondly, you can mix NESTOR and not-NESTOR metadata catalogs which allows for seamlessly integration of different visions of the managed *digital objects* within the same digital archive. This opens up the possibility of exploiting the whole breadth of methodologies and tools available in the DL field with the archives.

Future work will concern the formal definition of creation, deletion, update, and search operations on digital archives via NESTOR. This, in turn, will open up the possibility to further extend the 5S model. Indeed, according to it, a minimal digital library has to offer, at least, indexing, searching, and browsing services [4, p. 299].

Acknowledgments

CULTURA[3] (Grant agreement no. 269973) and the PROMISE network of excellence[4] (Contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

References

1. A. Agosti, N. Ferro, and G. Silvello. The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures. In *Proceedings of the 18th Italian Symposium on Advanced Database Systems*, pages 242–253. Società Editrice Esculapio, Bologna, Italy, 2010.
2. N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In B. Christensen-Dalsgaard et al., editor, *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, pages 268–279. Lecture Notes in Computer Science (LNCS) 5173, Springer, Heidelberg, Germany, 2008.
3. A. J. Gilliland-Swetland. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment.* Council on Library and Information Resources, Washington, DC, USA, 2000.
4. M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. *ACM Transactions on Information Systems (TOIS)*, 22(2):270–312, April 2004.
5. International Council on Archives. ISAD(G): General International Standard Archival Description, 2nd edition. Ottawa: International Council on Archives, March 1999.

[3] `http://www.cultura-strep.eu/`

[4] `http://www.promise-noe.eu/`