# "What is a Good Digital Library?" – A Quality Model for Digital Libraries

Marcos André Gonçalves [a] Bárbara L. Moreira [a] Edward A. Fox [b]
Layne T. Watson [b]

[a] *Department of Computer Science, Federal University of Minas Gerais, 31270-901 Belo Horizonte MG Brazil*

[b] *Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA*

**Abstract**

In this article, we elaborate on the meaning of quality in digital libraries (DLs) by proposing a model that is deeply grounded in a formal framework for digital libraries: 5S (Streams, Structures, Spaces, Scenarios, and Societies). For each major DL concept in the framework we formally define a number of dimensions of quality and propose a set of numerical indicators for those quality dimensions. In particular, we consider key concepts of a minimal DL: catalog, collection, digital object, metadata specification, repository, and services. Regarding quality dimensions, we consider: accessibility, accuracy, completeness, composability, conformance, consistency, effectiveness, efficiency, extensibility, pertinence, preservability, relevance, reliability, reusability, significance, similarity, and timeliness. Regarding measurement, we consider characteristics like: response time (with regard to efficiency), cost of migration (with respect to preservability), and number of service failures (to assess reliability). For some key DL concepts, the (quality dimension, numerical indicator) pairs are illustrated through their application to a number of "real-world" digital libraries. We also discuss connections between the proposed dimensions of DL quality and an expanded version of a workshop's consensus view of the life cycle of information in digital libraries. Such connections can be used to determine when and where quality issues can be measured, assessed, and improved — as well as how possible quality problems can be prevented, detected, and eliminated.

## 1 Introduction

What is a good digital library? As was pointed out in (Fuhr et al., 2001), the answer to this question depends on whom you ask. Many consider that what differentiates a good DL from a not so good one is the quality of its services and content. In previous work, we have sought to formally elaborate the notion of digital libraries using the 5S framework (Gonçalves et al., 2004). Since one of the main goals of that

work with 5S was to try to answer (at least partially) the question "What is a digital library?" our hypothesis in this article is that further development of the theory will allow us to define critical dimensions and indicators of DL quality. In contrast to its physical counterpart, the "digital" nature of digital libraries allows automatic assessment and enforcement of those quality properties, thereby supporting prevention and elimination of quality problems. 5S gives a standard terminology to discuss these issues in a common framework. Moreover, the formal nature of our DL theory allows us to add precision as we define specific DL quality dimensions and corresponding numeric indicators.

In this article, we will follow the standard terminology used in the *social sciences* (Babbie, 1990). We will use the term *composite quality indicator* [1] (or in short *quality indicator*) to refer to the proposed quantities instead of the stronger term *quality measure*. Only after one has a number of indicators, and they are validated [2] and tested for reliability [3], can they be composed into reliable "measures". Despite partial tests of validity (for example, through focus groups [4]) the proposed quality indicators do not qualify as measures yet. Also, it should be stressed that the proposed quantities are only approximations of or give quantified indication of a quality dimension. They should not be interpreted as a complete specification of a quality dimension, since more factors/variables could be relevant than are specified here. We will, however, reserve the right to use the term "measure" when talking about standard measures that have long been used by the CS / LIS communities. The distinction should be clear in context.

This article is organized as follows. Section 2 provides background and context necessary to understand the remainder of the article. Sections 3 through 6 present all the dimensions of quality, the proposed indicators, and their applications to key DL concepts. Section 7 deals with the connections between the proposed dimensions and Borgman et al.'s Information Life Cycle (Borgman, 1996). Section 8 shows the evaluation of the proposed quality model with a focus group. Section 9 covers related work and Section 10 concludes the article.

---

[1] An indicator composed of two or more simpler indicators or variables.

[2] According to (Babbie, 1990), validity refers to the extent to which a specific measurement provides data that relate to commonly accepted meanings of a particular concept. There are numerous yardsticks for determining validity: face validity, criterion-validity, content validity, and construct validity.

[3] Also according to (Babbie, 1990), reliability refers to the likelihood that a given measurement procedure will yield the same description of a given phenomena if that measurement is repeated.

[4] A type of face validity.

## 2 Background and Context

In this section, we summarize the 5S theory from (Gonçalves et al., 2004). Here we take a minimalist approach, i.e., we define, according to our analysis, the minimum set of concepts required for a system to be considered a digital library. Accordingly, let:

- $Streams$ be a set of streams, which are sequences of arbitrary types (e.g., bits, characters, pixels, frames);
- $Structs$ be a set of structures, which are tuples, $(G, L, F)$, where $G = (V, E)$ is a directed graph and $F : (V \cup E) \rightarrow L$ is a labeling function;
- $Sps$ be a set of spaces each of which can be a measurable, measure, probability, topological, metric, or vector space.
- $Scs = \{sc_1, sc_2, \ldots, sc_d\}$ is a set of scenarios where each $sc_k = \langle e_{1k}(\{p_{1k}\}), e_{2k}(\{p_{2k}\}), \ldots, e_{d_{kk}}(\{p_{d_{kk}}\})\rangle$ is a sequence of events tha t
  also can have a number of parameters $p_{ik}$. Events represent changes in computational states; parameters represent specific variables defining a state and their respective values.
- $St^2$ be a set of functions $\Psi : V \times Streams \rightarrow (N \times N)$ that associate nodes of a structure with a pair of natural numbers $(a, b)$ corresponding to a segment of a stream.
- $Coll = \{C_1, C_2, \ldots, C_f\}$ be a set of DL collections where each DL collection $C_k = \{do_{1k}, do_{2k}, \ldots, do_{f_k k}\}$ is a set of digital objects. Each digital object $do_k = (h_k, Stm_{1k}, Stt_{2k}, \Omega_k)$ is a tuple where $Stm_{1k} \subseteq Streams$, $Stt_{2k} \subseteq Structs$, $\Omega_k \subseteq St^2$, and $h_k$ is a handle which represents a unique identifier for the object.
- $Cat = \{DM_{C_1}, DM_{C_2}, \ldots, DM_{C_f}\}$ be a set of metadata catalogs for $Coll$ where each metadata catalog $DM_{C_k} = \{(h, mss_{hk})\}$, and $mss_{hk} = \{ms_{hk1}, ms_{hk2}, \ldots, ms_{hkn_{hk}}\}$ is a set of descriptive metadata specifications. Each descriptive metadata specification $ms_{hki}$ is a structure with atomic values (e.g., numbers, dates, strings) associated with nodes.
- Repository $R = (\{C_i\}_{i=1}^{f}, \{get, store, delete\})$ be a set of collections along with operations to manipulate them (see (Gonçalves et al., 2004) for details on the semantics of these operations).
- $Serv = \{Se_1, Se_2, \ldots, Se_s\}$ be a set of services where each service $Se_k = \{sc_{1k}, \ldots, sc_{s_k k}\}$ is described by a set of related scenarios. Any digital library should contain at least services for browsing, indexing, and searching.
- $Soc = (Comm, S)$ where $Comm$ is a set of communities and $S$ is a set of relationships among communities. $SM = \{sm_1, sm_2, \ldots, sm_j\}$ and $Ac = \{ac_1, ac_2, \ldots, ac_r\}$ are two such communities, where the former is a set of service managers responsible for running DL services and the latter is a set of actors that use those services [5]. Being basically an electronic entity, a member $sm_k$ of SM

[5] In this paper we will focus only on the relationships between and among actors and service managers that correspond to interactions mediated by the DL. We will not focus