Extending the 5S Digital Library (DL) Framework: From a Minimal DL towards a DL Reference Model

Uma Murthy
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
umurthy@vt.edu

Douglas Gorton
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
dogorton@vt.edu

Ricardo Torres
Institute of Computing
State University of Campinas
13084-851 Campinas, SP, Brazil
rtorres@ic.unicamp.br

Marcos André Gonçalves
Dept. of Computer Science
Federal University of Minas Gerais
Belo Horizonte, M.G., Brazil
mgoncalv@dcc.ufmg.br

Edward Fox
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
fox@vt.edu

Lois Delcambre
Dept. of Computer Science
Portland State University
Portland, OR 97207, USA
lmd@cs.pdx.edu

ABSTRACT

In this paper, we describe ongoing research in three DL projects that build upon a common foundation – the 5S DL framework. In each project, we extend the 5S framework to provide specifications for a particular type of DL service and/or system – finally, moving towards a DL reference model. In the first project, we are working on formalizing content-based image retrieval services in a DL. In the second project, we are developing specifications for a superimposed information-supported DL (combining annotation, hypertext, and knowledge management technologies). In the third effort, we have used the 5S framework to generate a practical DL system based on the DSpace software.

1. INTRODUCTION

DLs are immensely complex systems which allow information to be stored in an intelligent, usable, and easily retrievable fashion. In order to address the complexity of DLs, Gonçalves, et. al. proposed the 5S framework [8], where they defined a "core" or a "minimal" DL, i.e., the minimal set of components (a metamodel[1]) that make a DL, without which a system/application cannot be considered a DL. According to the framework, the nature of DLs can be described using the 5S's – Streams, Structures, Spaces, Scenarios, and Societies. Together these abstractions provide a formal foundation to define, relate, and unify concepts – among others, of digital objects, metadata, collections, and services – required to formalize and elucidate DLs. A reference model may be considered to be a structure or conceptual framework, which allows the modules of a system to be described and used in a consistent manner. Early versions of the DL reference model, as defined by the DELOS group, seemed to be aiming towards a comprehensive (maximal) representation of a DL, a DL system, and a DL management system [4]. The aim of this model is to facilitate the integration of research and to propose better ways of developing appropriate DL systems/applications.

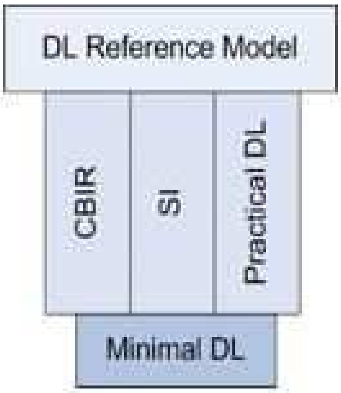

Figure 1: From a minimal DL to a DL reference model.

In this paper, we address three extensions of the 5S framework, going from a minimal DL (as described by the 5S framework) towards a (maximal or comprehensive) DL reference model. Figure 1 depicts this idea, where we consider a minimal DL as the foundation of various extensions, which serve as a base for a DL reference model. In the first extension, we are working on formalizing content-based image retrieval (CBIR) services in a DL (shown as CBIR). Clearly,

[1] *Metamodeling* is the construction of a collection of "concepts" (things, terms, etc.) within a certain domain. A model is an abstraction of phenomena in the real world, and a metamodel is yet another abstraction, highlighting properties of the model itself (from `http://www.wikipedia.org/`).

First International Workshop on "Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007
.

adding images to a DL is important, and since searching is a key service, CBIR services need to be supported. From the earliest days with many DL systems (such as electronic theses and dissertations), annotation was on top of the list of features to add. Also, given the importance of hypertext, having more specificity in hypertext, thus enabling working with information at sub-document granularities, seems to be of value. These ideas relate to our second extension, where we are developing a metamodel for a superimposed information-supported DL (combining specific features of annotation, hypertext, and knowledge management technologies, shown as SI). Finally, our third extension deals with DL generation based on DL software, such as DSpace in this case (shown as Practical DL). This is important because it helps to examine practical DL software functionality and architecture in the context of a formal DL specification, such as the 5S framework.

2. 5S FRAMEWORK

Recognizing the difficulties in understanding, defining, describing, and modeling digital libraries (DLs), Gonçalves, et al. have proposed and formalized the 5S (Streams, Structures, Spaces, Scenarios, and Societies) framework of DLs [8]. 5S provides a formal framework to capture the complexities of DLs. The definitions in [8] unambiguously specify many key characteristics and behaviors of DLs. This also enables automatic mapping from 5S constructs to actual implementations as well as the study of qualitative properties of these constructs (e.g., completeness, consistency) [6]. In this section, we summarize the 5S theory from [8]. Here we take a minimalist approach, i.e., we describe briefly, according to our analysis, the minimum set of concepts required for a system to be considered a digital library. ***Streams*** are sequences of arbitrary types (e.g., bits, characters, pixels, frames) and may be static or dynamic (such as audio and video). Streams describe properties of DL content such as encoding and language for textual material or particular forms of multimedia data. A ***structure*** specifies the way in which parts of a whole are arranged or organized. In DLs, structures can represent hypertexts, taxonomies, system connections, user relationships, and containment– to cite a few. A ***space*** is a set of objects together with operations on those objects that obey certain constraints. Spaces define logical and presentational views of several DL components, and can be of type measurable, measure, probability, topological, metric, or vector space. A ***scenario*** is a sequence of events that also can have a number of parameters. Events represent changes in computational states; parameters represent specific variables defining a state and their respective values. Scenarios detail the behavior of DL services. A ***society*** is "a set of entities and the relationships between them" and can include both human users of a system as well as automatic software entities which have a certain role in system operation. These 5Ss, along with fundamental set theoretic definitions, are used to define other DL constructs such as digital objects, metadata specification, collection, repository, and services.

Figure 2 shows concepts in the metamodel for a minimal DL using the 5S framework. For detailed formal definitions of the 5Ss and other DL constructs leading to the definition of a minimal DL, the reader is pointed to [6, 8]. The arrows in the figure indicate that some concepts are used in the definition of other concepts. For example, digital objects are composed of streams and structures. This representation is used in the metamodel figures that follow henceforth in sections 3, 4 and 5. Also, the extension figures in these sections have been drawn with the perspective of showing what needs to be added to the minimal DL. So, all DL concepts defined in the minimal DL (as mentioned in [8]) should be assumed to be in a DL that incorporates the extension.

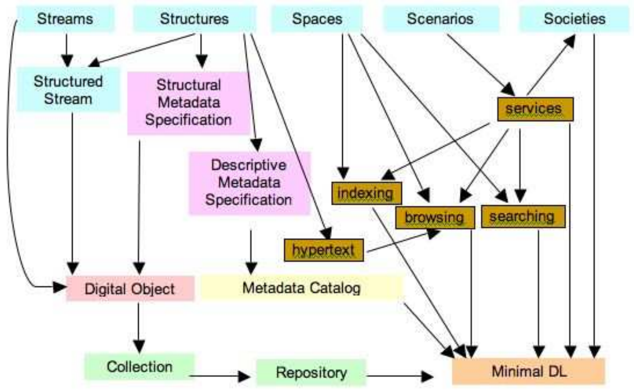

Figure 2: A minimal DL in the 5S framework.

3. CBIR SERVICES IN A DL

Technological improvements in image acquisition and the decreasing cost of storage devices have supported the dissemination of large image collections, supported by efficient retrieval services. One of the most common approaches involves *Content-Based Image Retrieval (CBIR) systems* [15, 17]. Basically, these systems try to retrieve images similar to a user-defined specification or pattern (e.g., shape sketch, image example). Their goal is to support image retrieval based on *content* properties (e.g., shape, color, or texture), usually encoded into *feature vectors*. One of the main advantages of CBIR is the possibility of an automatic retrieval process, avoiding the work of assigning keywords, which usually requires very laborious and time-consuming prior annotation of images.

Various Digital Libraries (DLs) support services based on image content [3, 5, 10, 14, 18, 19, 20, 21]. However, these systems are often designed and implemented without taking advantage of formal methods and frameworks. In this context, a research initiative is being conducted aiming to extend the 5S DL formal framework [8] for describing services based on image content description. The main contribution of this research is the proposal of several constructs that extend the 5S framework to handle image content descriptions and related services. These constructs can aid understanding of content-based image retrieval concepts as they apply to DLs. They also can guide the design and implementation of new DL services based on image content.

Figure 3 presents the proposed concepts based on the 5S framework to handle image content descriptions and related digital library services. A typical DL service based on image content information requires the construction of *image descriptors*, which are characterized by: (i) an *extraction algorithm* to encode image features into feature vectors; and (ii) a *similarity measure* to compare two images based on the distance between the corresponding feature vectors. The similarity measure is a *matching function*, which gives the

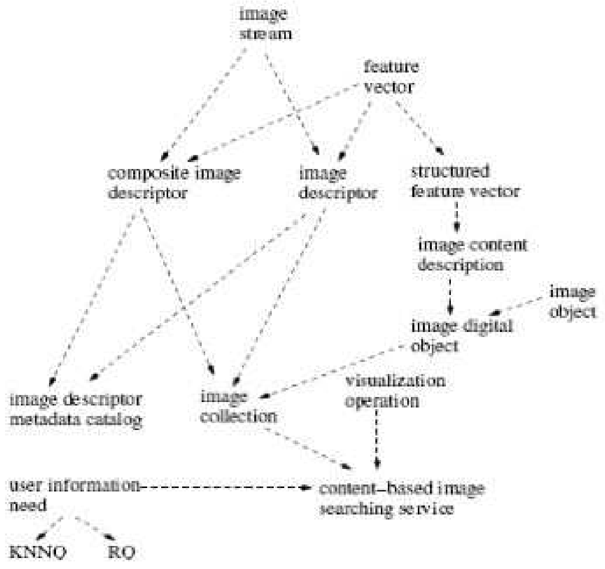
image stream
feature vector
composite image descriptor
image descriptor
structured feature vector
image content description
image object
image digital object
image descriptor metadata catalog
image collection
visualization operation
user information need
content–based image searching service
KNNQ
RQ

Figure 3: Formalizing CBIR services in a DL using the 5S framework.

degree of similarity for a given pair of images represented by their feature vectors, often defined as an inverse function of the distance (e.g., Euclidean), that is, the larger the distance value, the less similar the images. Structures can be applied to feature vectors for storage purposes (*structured feature vector*) and *image digital object* is defined by extending the original 5S digital object concept by considering image content descriptions. Two typical searching services based on image content can be usually performed: K-nearest neighbor query (KNNQ) and range query (RQ). In a KNNQ, the user specifies the number k of images to be retrieved closest to the query pattern. In a RQ, the user defines a search radius r and wants to retrieve all database images whose distance to the query pattern is less than r.

4. AN SI-SUPPORTED DL

For digital libraries (DLs) to fully support domains such as education there is a need for capabilities that go beyond information seeking-related services. DL users need, but get very little help with:

- Selecting and annotating multimedia information at varying document granularities – parts of a document, to a complete document, to multiple documents
- Linking new content with existing content, at varying document granularities
- Organizing/arranging annotated information.
- Sharing and reusing of new information (annotations, structures, etc) and associated existing information
- Finding and re-finding new information (annotations, structures, etc) and associated existing information through searching/browsing/visualization

An example of such use could be by a Biology professor, who is preparing for a class on the brain. Most of her class material comes from existing (multimedia) resources. For a particular topic, she wants to be able to work with pieces of information in various documents and prepare lecture notes, course materials, presentations, etc. Then, she wants to be able to share all this information with her students and with other faculty, who may have their own representation of the same information.

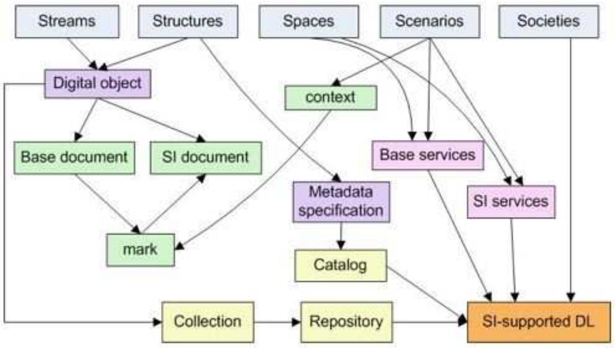

Figure 4: An SI-supported DL using the 5S framework.

Existing DLs such as [2, 16] facilitate some of these tasks; however, they provide limited support for working with heterogeneous multimedia formats and/or for working with information at varying document granularities while retaining original information context. We are working towards the development of a Superimposed Information-Supported Digital Library (henceforth referred to as SI-DL), which will bring together *superimposed information* along with traditional DL services that operate in context (of a domain such as education). We believe this will help in building a system with functionality to support annotation, linking, knowledge management and, sharing and reuse of information in tasks such as those mentioned above. Superimposed information (SI) refers to new information laid over existing information [11]. It is supplemental information created to reference, highlight, and extend information present elsewhere. Examples cover a variety of new interpretations, including annotations, tags, citations, indexes, concept maps, multimedia presentations, etc. The focus of SI research is to enable working with sub-document information, such that a user may (a) deal with information at varying document granularity, and (b) select or work with information elements at sub-document level while retaining the original context (by referencing, not replicating, information).

Beginning with development of scenarios and applications (such as [12, 13, 14]), literature review, and brainstorming, we have come up with a preliminary set of specifications for an SI-DL. To ground our work on a firm theoretical foundation, we are extending the 5S framework for DLs to formally define essential units in an SI-DL, resulting in an SI-DL metamodel. These constructs will not only aid in a deeper understanding of SI and related concepts, but also will serve as building blocks for defining various possibilities of an SI-DL.

Figure 4 shows our preliminary work in identifying important SI concepts and their relation to 5S constructs. At the core of an SI system, is a *mark* – an abstraction that specifies an addressable/reference-able region or sub-document, in existing multimedia information of heterogeneous formats.

Marks connect base documents and SI documents. A *base document* is information already existing in the digital library and marks are created in a base document. Marks are used in *SI documents*, which may be constructed by organizing marks in a specific schema/structure. *Context* refers to information and conditions surrounding creation and use of SI including mark creation context, usage context, and context associated with software dealing with base documents and SI documents. Apart from existing *base services* (such as search, browsing and indexing), an SI-DL has *SI services*, which support creation, use and management of marks, context and SI. Finally, creators, viewers, and users of SI form societies that will interact with SI.

5. A PRACTICAL DL

In today's ever-changing world of technology and information, a growing number of organizations and universities seek to store digital documents in an online, easily accessible manner. DLs provide the medium for the online storage and dissemination of such documents and many open source and commercial products are available that help users accomplish that task. While DL software packages enable a broader adoption of DLs, there is still a certain amount of configuration, customization, and data ingestion that must occur in such systems before they are truly optimally usable and set up to serve as many of the institution's needs as allowable. The generation of DLs attempts to abstract some of these processes into a simpler, clearer task where the nature of the desired digital library is described and the generator handles those details with regard to configuration, customization, generation of pertinent code, etc. The intent is to automate these tasks in a way that the DL designer has an appreciation and understanding of the repository to be created but does not need to worry about the underlying technological layer as would be needed if the DL were created manually.

In order to ease the process of creating DLs, we have created a XML-based specification model that describes the nature of possible DLs in MIT and HP Labs' DSpace DL software [1]. We base our work on DL specifications on Fox and Gonçalves' work with the 5S Framework for Digital Libraries [8] and its domain specific digital library declaration language, 5SL [7]. While the original work with DL specification with 5SL was complete, it was more suited to theoretically describe DL systems. In this work we move to a more practical DL metamodel and apply the aspects of 5S and 5SL to describe the nature, structure, and functionality of a modern DL system such as DSpace.

Figure 5 represents the essential components of this metamodel. In order to continue to use a 5S driven organization and separation of the concerns of a digital library, it is necessary to examine the DSpace functionality and architecture in the context of the 5 S's. Thus, we decompose the functionality, structure, and services of DSpace into the aspects that the 5S framework suggests. Because DSpace is a mature, open source software project that has much built-in capabilities as well as customizations via source code and other avenues, our work focuses only on the most commonly used aspects of DSpace. For example, the main DSpace organizational components are Collections and Communities, where Communities are sets of Collections with documents of similar content and subject matter, which we apply to the original hierarchical 5SL constructs of Collections and CollectionSets. Each XML element representing either of these aspects of a DSpace DL also has sub-elements which describe metadata characteristics of each, such as a name, description, and textual components to be used in interfaces. Users that are desired in a DSpace DL are described in a Society sub-model, split into collections of ÔManagers' for administrative users and ÔActors' for regular users. Each type of user requires a few defined metadata elements needed in DSpace such as a password, name, and phone number. Groups of users are also similarly defined.

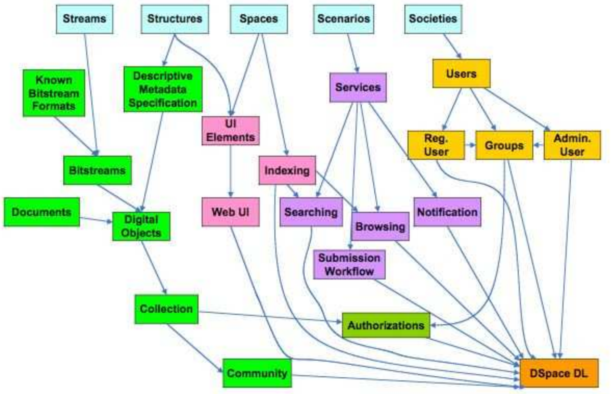

Figure 5: A practical (DSpace) DL using the 5S framework.

This work with DSpace generation provides a good proof of concept for applying past work with DL specification and generation to a widely used repository system but there is still much work to be done with DL specification and generation in general. Choices needed to be made to decide which DSpace functionalities were supported for specification and generation, and due to that some functions were unable to be created programmatically by the generator. Much additional work can be done to provide a more comprehensive and all encompassing specification and generation ability for DSpace. Similarly, there are many DL packages out there that have different strengths and are well suited for different applications–the eventual move toward more generalized ways of specifying and generation DL systems would lead to a more streamlined consistent installation and generation path for all these systems. For details on this work, the reader is pointed to [9].

6. TOWARDS A DL REFERENCE MODEL

We have described three different efforts in progress, which build upon a common foundation – the 5S minimal DL framework. Of course there are other extensions also needed. However, one needs to start somewhere and certainly these extensions serve as distinct and valuable starting points. We consider the development of the aforementioned extensions as a step towards understanding, comparing and combining results achieved in different areas of DL development – thus, serving as a base for the development of a DL reference model. Just as there needs to be eventual movement towards a broadly applicable model for DL specification, a more framework oriented approach for the generation of DLs based on specifications is also a direction that would allow for easier, more consistent DL generation. Figure 6 shows such a generation process. We begin by defining a meta-

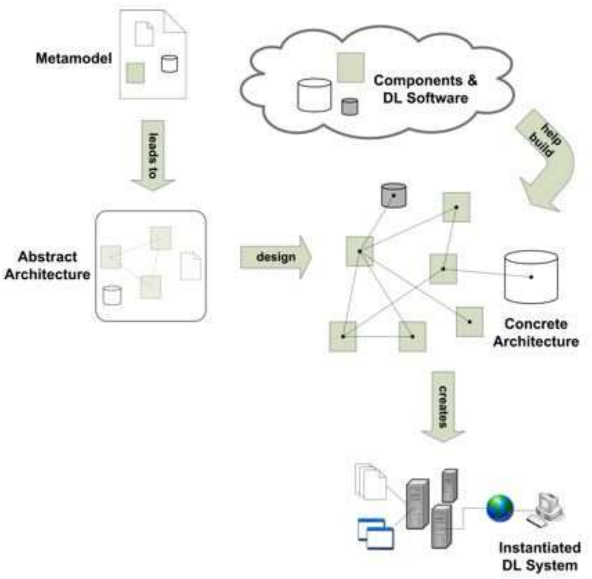

Figure 6: A DL generation process.

model of constructs, or building blocks for the specific DL we want to generate. Specific instances of this metamodel may be derived that represent a user's desired DL system and make up an abstract DL architecture. Based on the declared DL and available software components (and systems, such as DSpace), a concrete architecture may be created for that DL, which finally may be built into a DL system.

Acknowledgments

Thanks go to sponsors of our work, which include: AFOSR (grant F49620-02-1-0090), AOL, IMLS, NSF (grants DUE-0121679, DUE-0435059, IIS-0325579, IIS-0535057) and CNPq project 5S-VQ (grant MCT/CNPq/CT-INFO 551013/2005-2), FAPESP, FAEPEX, Microsoft Escience project, and the Microsoft tablet PC grant. Thanks also go to those in Virginia Tech's Digital Library Research Laboratory, and to all those involved in our various projects.

7. REFERENCES

[1] DSpace Federation, http://dspace.org/.

[2] Agosti, M., Albrechtsen, H., Ferro, N., Frommholz, I., Hansen, P., Orio, N., Panizzi, E., Pejtersen, A.M. and Thiel, U. DiLAS: a Digital Library Annotation Service. Presented at the International Workshop on Annotation for Collaboration, Paris, Framce, 2005.

[3] Bergman, L.D., Castelli, V. and Li, C.-S. Progressive Content-Based Retrieval from Satellite Image Archives. D-Lib Magazine, 3 (10).

[4] Candela, L., Castelli, D., Ioannidis, Y., Koutrika, G., Pagano, P., Ross, S., Schek, H.-J. and Schuldt, H. A Reference Model for Digital Library Management Systems. http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid=#docs.

[5] French, J.C., Chapin, A.C. and Martin, W.N. An application of multiple viewpoints to content-based image retrieval. In proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries, (Houston, Texas, 2003), IEEE Computer Society, 128-130.

[6] Gonçalves, M. Streams, Structures, Spaces,Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications. PhD Dissertation, Computer Science, Virginia Tech, Blacksburg, 2004. http://scholar.lib.vt.edu/theses/available/etd-12052004-135923/.

[7] Gonçalves, M. and Fox, E.A. 5SL: a language for declarative specification and generation of digital libraries. In proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, (Portland, Oregon, USA, 2002), ACM Press, 263-272.

[8] Gonçalves, M.A., Fox, E.A., Watson, L.T. and Kipp, N.A. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM TOIS, 22 (2). 270-312.

[9] Gorton, D. Practical Digital Library Generation into DSpace with the 5S Framework. Master's thesis, Computer Science, Virginia Tech, Blacksburg, 2007. http://scholar.lib.vt.edu/theses/available/etd-04252007-161736/.

[10] Hong, J.S., Chen, H.Y. and Hsiang, J. A Digital Museum of Taiwanese Butterflies. In proceedings of the Fifth ACM Conference on digital Libraries, (San Antonio, Texas, United States, 2000), 260-261.

[11] Maier, D. and Delcambre, L. Superimposed Information for the Internet. in WebDB Workshop, (1999), 1-9.

[12] Murthy, U., Ahuja, K., Murthy, S. and Fox, E.A. SIMPEL: a superimposed multimedia presentation editor and player. In proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. 377.

[13] Murthy, U., Richardson, R., Fox, E.A. and Delcambre, L. Enahcing Concept Mapping Tools Below and Above to Facilitate the Use of Superimposed Information. In proceedings of the Second International Conference on Concept Mapping, (San Jose, Costa Rica, 2006).

[14] Murthy, U., Torres, R.d.S. and Fox, E.A. SIERRA: A Superimposed Application for Enhanced Image Description and Retrieval. Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries, 2006, 540-543, http://dx.doi.org/10.1007/11863878_63

[15] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (12). 1349-1380.

[16] Sumner, T., Ahmad, F., Bhushan, S., Gu, Q., Molina, F., Willard, S., Wright, M., Davis, L. and Janèe, G. Linking learning goals and educational resources through interactive concept map visualizations. International Journal on Digital Libraries, 5 (1). 18-24.

[17] Torres, R.d.S. and Falcão, A.X. Content-Based Image Retrieval: Theory and Applications. Revista de Informática Teórica e Aplicada, 13 (2). 161-185.

[18] Vemuri, N.S., Torres, R.d.S., Gonçalves, M.A.,

Fan, W. and Fox, E.A. A Content-Based Image Retrieval Service for Archaeology Collections. In proceedings of the European Conference on Digital Libraries, (Alicante, Espanha, 2006), 438-440.

[19] Wang, J.Z. and Du, Y. Scalable integrated region-based image retrieval using IRM and statistical clustering. In proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, (Roanoke, Virginia, USA, 2001), 268-277.

[20] Wang, Y., Makedon, F., Ford, J., Shen, L. and Goldin, D. Generating fuzzy semantic metadata describing spatial relations from images using the R-histogram. In proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries, (Tuscon, AZ, USA, 2004), 202-211.

[21] Zhu, B., Ramsey, M. and Chen, H. Creating a Large-Scale Content-Based Airphoto Image Digital Library. IEEE Transactions on Image Processing, 9 (1). 163-167.