Available online at www.sciencedirect.com
ScienceDirect
Information Systems 33 (2008) 699–723
www.elsevier.com/locate/infosys

Integration of complex archeology digital libraries: An ETANA-DL experience

Rao Shen, Naga Srinivas Vemuri, Weiguo Fan*, Edward A. Fox

Digital Library Research Laboratory, Virginia Tech, Blacksburg, VA 24061, USA

Abstract

In this paper, we formalize the digital library (DL) integration problem and propose an overall approach based on the 5S (streams, structures, spaces, scenarios, and societies) framework. We then apply that framework to integrate domain-specific (archeological) DLs, illustrating our solutions for key problems in DL integration. An integrated Archeological DL, ETANA-DL, is used as a case study to justify and evaluate our DL integration approach. More specifically, we develop a minimal metamodel for archeological DLs within the 5S theory. We implement the 5SSuite tool set to cover the process of union DL generation, including requirements gathering, conceptual modeling, rapid prototyping, and code generation. 5SSuite consists of 5SGraph, 5SGen, and SchemaMapper, each of which plays an important role in DL integration. We also propose an approach to integrated DLs based on the 5S formalism, which provides a systematic method to design and implement DL exploring services.

Keywords: Information integration; Digital library integration; Metadata; Integration; 5S; Schema mapping; Information visualization; Archaeology; Semantic interoperability; Exploring service; Multi-dimensional browsing

1. Introduction

1.1. The problem

Digital libraries (DLs) are transforming research, scholarship, and education at all levels. One of the intriguing aspects of DL research is that challenges exist at both the fundamental technology level and at the large-scale integration level. Over a decade of government and private funding of DL research projects has led to important results at the fundamental technology level. The successes in large-scale integration are arguably less evident. Even the notion of "DL integration" is ambiguous in the sense that different approaches and proposed solutions exist. Work on DL integration focuses to an extent on three issues [1]:

(1) *Distribution*: geographical spread;
(2) *Heterogeneity*: difference at both the technical level (e.g., hardware platform, operating system, programming language, etc.) and conceptual level (e.g., different understanding and modeling of the same real-world entity);
(3) *Autonomy*: the extent to which the components are self-sufficient, as opposed to being delegated a role only as components in a larger system.

By "DL integration", we mean hiding distribution and heterogeneity, while at the same time

*Corresponding author. Tel.: +1 540 231 6588.
E-mail addresses: rshen@vt.edu (R. Shen), nvemuri@vt.edu (N.S. Vemuri), wfan@vt.edu (W. Fan), fox@vt.edu (E.A. Fox).

0306-4379/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2008.02.006

enabling and making visible component autonomy (at least to some degree).

Many DLs belonging to different autonomous organizations were developed independently without plans to provide open and easy automated access to their data and functionality. The inability to seamlessly and transparently access knowledge across DLs is a major impediment to knowledge sharing. The goal of DL integration then is to utilize various autonomous DLs in concert to provide knowledge hidden in such island-DLs. The needs for DL integration are well known, and better known than the solutions [2].

Challenges to DL integration are a direct result of DL characteristics. DLs are complex information systems due to their inherently interdisciplinary nature, both with regard to application domains and technologies involved in building the systems. Concerning the latter, DL system implementations integrate findings from disciplines such as hypertext, information retrieval, multimedia services, database management, and human–computer interaction [3]. Hence, an integrative theory for DLs is needed. Gonçalves et al. [4] summarizes key early work on the 5S (streams, structures, spaces, scenarios, and societies) framework, and related efforts to construct such an integrative theory for DLs. The 5S framework allows us to define DLs rigorously and usefully. Streams are sequences of arbitrary items used to describe both static and dynamic (e.g., video) content. Structures can be viewed as labeled directed graphs, which impose organization. Spaces are sets with operations that obey certain constraints. Scenarios consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement. Societies are sets of entities and activities, and the relationships among them. 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 "digital libraries" [5].

DL integration can be done at different levels, e.g., information level and service level. Integrated information makes distributed collections of heterogeneous resources appear to be a single union. Integrated services provide users more comprehensive usage of DL resources through more coherent and easier to use interfaces that hide syntax and semantic differences in the DLs to be integrated.

Developing an infrastructure to address all perspectives of the DL integration problem is an ambitious task. While many efforts have looked into the DL integration problem, most developed their own approaches in an ad hoc and piecemeal fashion. Building upon our previous work on DL integration [6,7], we formalize in this paper the DL integration problem and propose an overall approach based on the 5S framework. We then apply our framework to integrate domain-specific (archeological) DLs, illustrating our approaches to key sub-problems (e.g, semantic interoperability) of DL integration. Thus we make contributions at both the methodological and practical levels (implementation and prototyping).

1.2. Hypothesis and research questions

We claim that the 5S framework provides effective solutions to DL integration. This hypothesis leads to the following research questions.

(1) Can we formally define the DL integration problem, using the 5S framework?
(2) Can the 5S framework guide integration of domain/discipline focused DLs (e.g., integrate systems for diverse archeological sites into a union archeological DL)? If yes, how? Specifically:

- How can we formally model such domain-specific DLs in the 5S framework?
- How can we integrate DL models into a union DL model?
- How can we use the union DL model to help design and implement high-quality integrated DLs?

1.3. Problem formalization and overall approach

Formalizing DL integration facilitates the development, comparison, and evaluation of solutions; makes clear to users what a solution means; and helps users evaluate the applicability of a solution. Furthermore, it allows us to leverage special-purpose techniques for the DL integration process. In this section, we first give a background to the 5S framework, based on which we formally define the DL integration problem. We then propose an overall approach and a toolkit for DL integration.