INTEGRATING DISTRIBUTED HETEROGENEOUS INFORMATION SOURCES FOR CULTURAL HERITAGE: THE DICE APPROACH

Vito Perrone
HOC.Labs
Politecnico di Milano
Milano, Italy
perrone@elet.polimi.it

Chiara Bramani
HOC.Labs
Politecnico di Milano
Milano, Italy
cbramani@como.polimi.it

Sebastiano Colazzo
HOC.Labs
Politecnico di Milano
Milano, Italy
colazzo@elet.polimi.it

Abstract

The aim of DICE is the improvement of communication between culture-providers (institutions, museums, etc.) and different categories of users (scholars, publishers, cultural-tourism promoters, etc.), trying to "translate" into a common vocabulary the patrimony of content held by each source of information. The result is a shared virtual space, allowing access to different sources of information, integrated in a peer-to-peer philosophy, in order to create an effective working environment for professional users.

DICE provides a technical infrastructure, a software platform, an organizational and cultural model, with a flexible and scalable approach that allows the collection of information from all the actors involved.

INTRODUCTION

The Cultural Heritage plays a relevant role for every country both from cultural and economic points of view since it represents a considerable attractive for tourists.

According to the Italian law as it stands [1], a "cultural asset" is anything composing the historical, artistic, monumental, archaeological, demo-ethno-anthropological and librarian that bears witness the Italian culture. Cultural assets in Italy are an immense resource that only nowadays is getting the real importance becoming a mass phenomenon. Italy hosts one of the largest collections of cultural heritage world wide, scattered all over its territory and belonging to different subjects (the Government, the Church, private owners, etc.). This situation has determined the creation of a host of different sources of information, some of them public, others private, all of them "uncoordinated". The current scenario reveals an enormous proliferation of information sources that are local, scattered, heterogeneous, little known, barely accessible, scarcely reused, and hardly reusable. If, today, a professional wants to collect all the information about a possible item of interest, this means a large waste of time (and money) looking around at different sources.

Over the last few years the improvement of the new technologies is offering new opportunities for accessing this immense resource and the central government is promoting initiatives aiming at taking advantage of the opportunities offered by the new technologies for this crucial economical sector. The deep understanding of the cultural patrimony requires that its informative sources are studied in an integrated perspective, in that researchers or professionals should be able to access information or discovering new ones even if belonging to different information sources geographically spread. A valuable integration system should exploit the integrate environment by enabling to search through all information elements and to compare them but it should also enable the creation of new forms of content.

This scenario is hardly feasible in the current practice, for technical, organizational, cultural reasons. A review of the informational "procedures" [2] proves how much this field

has been characterized by projects developed with "proprietary" approaches, hardly portable both from the technological and methodological side.

Some research projects [3,4,5,6,7,9] have attempted at integrating heterogeneous information sources but these approaches aim at integrating the information resources of a single information producer, and not at building "large scale" integration, which is the aim of the DICE project. Each subject tends to build his own model of information representation [4], developing his/her terminological dictionaries and elaborations to address his/her own needs, often imposing a proprietary data structure for digital repositories.

The Italian Ministry of Cultural Heritage provides, through the ICCD [10], a sort of "schema" to represent cultural heritage data, which is however mandatory only for the activities performed by Soprintendenze (Directorates). At international level, existing researches, such as the Dublin Core Metadata Initiative [8] , have focused on the definition of meta-representation standards or "meta-data models" [11]. These provide a "lingua franca" to describe information sources at conceptual level, so that data differing in semantics, medium, structure, technical format, may have the same representation at the meta-level based on common terms and rules. Thus, metadata models offer a shared conceptual base for the mediation and exchange of information among cultural institutions. These models have limited efficacy in the Cultural heritage field, as it is mainly aimed to standardize the information sources.

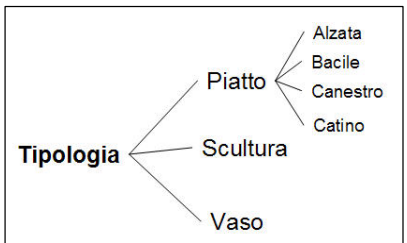
DICE (Distributed Infrastructure for Cultural hEritage) is a project, co-financed by the Ministry of the University and scientific Research, involving several companies and leading Universities (Politecnico di Milano and Scuola Normale di Pisa). The main goal of DICE is to demonstrate the possibility of integrating different sources of information, in order to create an effective working environment for professional users: researchers, scholars, cultural-tourism promoters, etc. A DICE demonstrator was implemented in March 2004, integrating more than 20 different sources both for "Archaeology in Campania" and for "Ceramic in Campania", holding more than 3,000 pieces of information. Information providers are leading institutions, researchers, publishers, etc, while users are scientists, researchers, publishers, writers, tourism promoters, etc. In the following we will describe in detail the DICE approach, the way the demonstrator was developed and its results.

THE DICE APPROACH

From the organizational point of view, DICE has a scalable bottom-up approach. DICE encourages the development of cultural communities, integrating different information sources that participate in it by sharing their own information around a specific topic. The technical basis of DICE is a "peer-to-peer" infrastructure: each owner of information, participating in a DICE community, makes its content available, but still retaining full control over it from its actual location: the user perceives the "universe of information" as a seamless hypermedia, with the different items interconnected in a network.

The very distinctive feature of DICE concerns the content' organization. In DICE each possible item of information is called *information unit*: it belongs to a single information source, it can be a museum's card, an image with a description, a paragraph of a catalogue describing a work of art, etc. What glues all the sources together is a common cultural model which answers to questions like: "How should the information be represented?", "Which are the possible association between cultural assets?", "How should the information be organized and classified so that users can intuitively access them?". The cultural model of a DICE community represents a shared local ontology agreed by all the community members.

The Cultural Model



The cultural model consists of set of taxonomies, a set of profile schemas, a set of content schemas and a set of semantic association types.

A taxonomy, see Figure 1, is a structured vocabulary apt at describing a specific domain (for example ceramic or archeology).

Figure 1. Taxonomy – "Tipologia"

Each information source has to "describe" through a "semantic profile schema" (a set of categories) and a "content schema" (a set of macro-paragraphs that describe the conceptual structure information). When a new information source joins to a DICE community, in order to support the integration with all others information sources and to allow a unified access to their information, both content and semantic profile schemas have to be chosen for representing and characterizing its own information. From the above it follows that an information unit (belonging to an information source) is made of a content, instance of the corresponding content schema, and a semantic profile, instance of the corresponding semantic profile schema. For example, if an information unit belonging to an information source, whose content and profile schema are represented in Table 1, describes a ceramic object, the "Technique", the "Location", the "Chronology" and the "Decorative subject" will be specified, choosing for each from the corresponding taxonomy. Similarly, the paragraphs of the content schema will be filled with the corresponding information.

	Category
Semantic Profile Schema	Technique Location Decorative subject Chronology
	Paragraph
Content Shema	Technical Data Historical Data Description Image

Table 1. "Semantic Profile Schema" and "Content Schema"

It is important to notice that the semantic profile is not conceived for user consumption but it is only used for setting links among different cultural assets that share some common characteristics and to allow the user to find them. On the other hand, the content is only for user consumption.

Moreover, separation between semantic profile and content of an information unit allows having two levels of integration, the content level and the visual level. At the content level, profiles allow the user to access information units, which come from different information sources, in an integrated fashion (not perceiving its fragmentation). At the visual level, on the other hand, a uniform representation of contents allows the user to improve his "user experience" with a considerably improving of the application usability and acceptability.

A semantic association type defines, from the user point of view, a possibility of navigation among information about cultural assets shared by community users. Examples of possible association types are:

- *Identity relationship*: information about a cultural asset that can be related to other information since they refer to the same cultural asset (e.g. they describe the same vase).
- *Physical relationships*: cultural assets may be related to other cultural assets by physical relationships, such as "it belongs to", or "it was found in", or "it is part of", etc.
- *Semantic relationship*: cultural assets can be related to other cultural assets by a variety of semantic relationships (e.g. they have the same decorative subject).

Another distinctive feature of our approach is that different taxonomies, content and profile schema, and association types are built according to different end users' profile: for example, an expert user in fact will be able to understand the meaning of technical words (such as "oinokoe", a kind of Greek pottery), whilst a lay user will not.

THE DEMONSTRATOR

A first DICE demonstrator has been implemented in March 2004, in order to demonstrate the validity of the DICE approach. The sources of the information integrated inside the demonstrator has been taken from two main cultural fields: archaeology and ceramic in particular in the area of Napoli and Campania (a region in the South of Italy). These sources of information (more than 40 on the whole) public or private (leading institutions, museums, researchers, publishers, etc) were mainly data bases, catalogues, books, etc... The possible users considered in the demonstrator were divided into two main categories, that is experts (public cultural administration, researchers, scholars, etc...) and professional users (such as tour operators, cultural tourism promoters and publishers). The result has been a web site where users can accede to the integrated information about cultural heritage easily and in an efficient way, obtaining correlations among the different information from the various sources.

Accessing the Information Units

DICE allows the users to identify the relevant units of information (independently from the source they belong to) with four possible paradigms:

1. *Predefined cultural access paths*: each community will develop collections of information units corresponding to a specific cultural purpose (e.g. "the riggiole in S.Chiara"). Cultural access paths can be defined from each user of the community and they are used to build indexes which help users to find their way around. The user can choose a path and see the information unit linked to it.

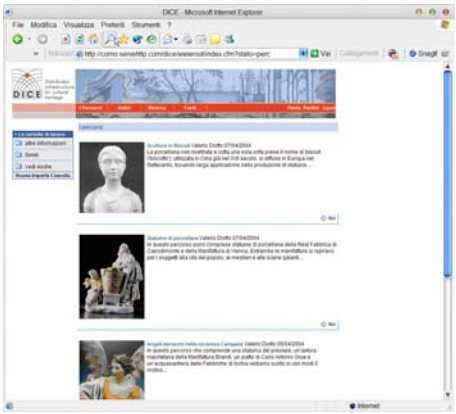


Figure 2. Predefined cultural access paths

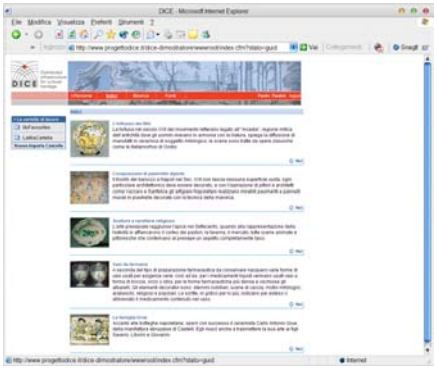
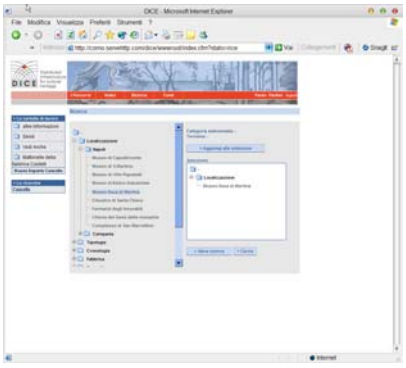


Figure 3. List of guided accesses

2. *Guided access*: a set of hierarchical indexes will partition the material in a way that is effective for the users. Many different schemes of indexes can be implemented (from some expert users) in order to facilitate access to the user. User can navigate towards information with the help of these guided accesses selecting given categories step by step.

3. *Advanced search:* using semantic profiles and predefined taxonomies users may search according to their own specific criteria. The result of the research is a list of information units that meet the research's criteria.



4. *Hypertext navigation*: starting from a specific information unit, say a IU describing a vase, the user may "navigate" to other units (likely) describing the same vase or to other units (likely) describing similar vases (similar for the subject, or the technique,...). These links are defined by specific programs which run in order to exploit specific situations on the basis of the profiles associated to each IU.

Figure 4. Advanced search: criteria selection

THE LOGICAL INFRASTRUCTURE

The design of the DICE infrastructure, enabling the information sharing among information sources and cultural heritage professionals, has been driven by a set of requirements, assumptions and constraints acquired in the early analysis of this domain in Italy. In particular, the key points surrounding the infrastructure design are summarized in the following list:

- *Information ownership*: each participant of a DICE community must keep the ownership of its information.
- *Joint community*: within a DICE community all the participant have the same power, that is, the community survival cannot be subordinated to the existence of a particular organization.
- *Cultural model agreement:* when a new information source joins a DICE community it must agree to the relative cultural model
- *Different kinds of participants' architectures:* due to the different kind of users DICE addresses to, the infrastructure should support the joining of various kinds of systems. In any case, it should be noticed that the access to a DICE community is restricted to known users who are all professionals.
- *Scalability:* The infrastructure should be easily scalable in terms of number of participants of a community.

The architectural solutions devised for meeting such requirements is represented in Figure 5 where possible configurations of participants are shown.

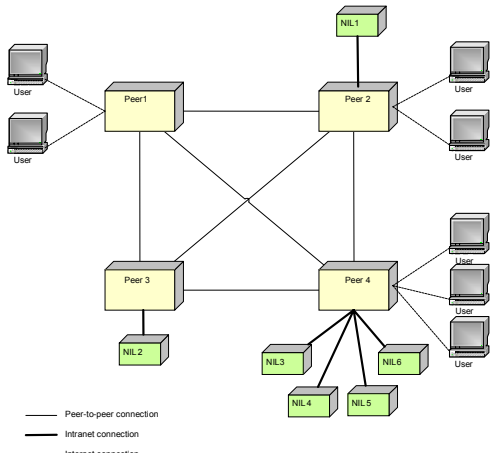


Figure 5: DICE' logical infrastructure

In general, DICE is a system composed by several nodes which offer services (like access to the network) and/or information to other nodes. Possible node configurations are: full peer (peer2 in Figure X that contains a NIL, that is an information source, and provides access to the DICE community to its users), data peer (peer3 in Figure ? where a peer is only equipped with hardware and software enabling information sharing), access provider (peer1 in Figure 5 that represents an organization which participate the DICE community for providing access to other users) and virtual NIL host (peer4 in Figure 4 that extends the access provider peer offering a host service for virtual NILs belonging to organizations which own interesting information sources but that cannot afford the hardware and software configuration needed for being a full peer).

FINAL CONSIDERATIONS

The goal of the DICE projects is the design and implementation of an innovative infrastructure enabling communities of professional users sharing and accessing the existing information concerning the cultural heritage in well-defined areas (a region, a zone in the territory) and about a set of specific topics ("Greek archaeology in Cuma", "Naple's ceramic", etc.). After two years, the main results achieved in the DICE project can be summarized in the design, implementation and usage of an innovative methodology for integrating very heterogeneous contents and in the design and implementation of a working platform. In the former, we have highlighted the need for integrating information from two perspectives. The first, the user perspective, provides a unique an intuitive environment for accessing very different contents. The second, the system perspective, fronts the problem of providing a common language for representing objects belonging to the same cultural field, without imposing any new data format. Splitting what users visually perceive from the information needed for running the system, we have achieved the fundamental result of keeping the respective terminology of each source unmodified even though integrated with others belonging to very different sources. From the technological point, the DICE infrastructure combines a simple management with a strong flexibility that allows several kinds of configurations available to meet different requirements of various kinds of participants.

ACKNOWLEDGMENTS

The work presented in this paper has been partially funded by the DICE research project. Authors are very grateful to all the partners of the project – EDA spa, eWorks spa, CEFRIEL and FORMA – for their valuable contribution in all the discussions about cultural and technological aspects. We also thank the many cultural experts who have cooperated to the approach definition and validation.

References

[1] Ministero per I beni e per le attività culturali, http://www.beniculturali.gov.it/

[2] A. D'Andrea, F. Niccolucci, 'L'archeologia computazionale in Italia: orientamenti, metodi e prospettive", Proceedings of I National Workshop Archeologia Computazionale, Napoli-Firenze 1999.

[3] M. Valenti, "La piattaforma GIS dello scavo. Filosofia di lavoro e provocazioni, modello dei dati e soluzione GIS.", Proceedings of I National Workshop Archeologia Computazionale, Napoli-Firenze 1999.

[4] A. Santoriello, F. Scelza, "Un sistema informativo archeologico: l'applicazione del Syslat a Fratte di Salerno", Proceedings of I National Workshop Archeologia Computazionale, Napoli-Firenze 1999.

[5] M. Crescioli, F. Niccolucci, C. Tonghini, G. Tannini, "PETRA: un sistema integrato per la gestione dei dati archeologici.", Proceedings of I National Workshop Archeologia Computazionale, Napoli-Firenze 1999.

[6] R. Francovich, V. Fronza, A. Nardini, M. Valenti, "Open Archeo: An Information System for Archaeological Data Management. Recent Developments and Future Aims", in proceeding of conference on Electronic Imaging k the Visual Arts, Italy, 2003

[7] L. Devile, S. Lissonett, "Dublin Core: The base for an indigenous culture environment.", In Proceedings Museum and the Web 2003, Charlotte (NC), USA 2003

[8] Dublin Core, http://dublincore.org.

[9] Canadian Heritage Information Network Project, http://www.chin.gc.ca/

[10] Central Institute for Cataloguing and Documentation, http://www.iccd.beniculturali.it

[11] ISO - International Organization for Standardization, http://www.iso.ch