An ontological approach to the integration of information for Cultural Heritage: the DICE project

Colazzo Sebastiano, Paolini Paolo, Vito Perrone.
HOC-Hypermedia Open Center, Department of Electronics and Information
Politecnico di Milano – Italy,

colazzo|paolini|perrone@elet.polimi.it

1. Introduction

Thanks to the explosion, happened over the last years, of data available to be accessed by networks (both Internet and intranet within a corporate), finding and integrating information coming from different sources has become a crucial factor for the enhancement of the informative assets of a corporate.

On the other hand, many tasks that should be performed by users of data based complex systems impose the interaction with a multiplicity of information sources.

Knowledge Management has developed to satisfy these needs; it includes all methodologies and technologies that allow managing both the knowledge and the key information in an organization.

The knowledge can be unspoken or outspoken. The unspoken knowledge, or implicit, belongs to the deep of each person and it depends from his (her) spirituality, from his (her) culture, from his (her) personality and from his (her) own individuality; it is a type of knowledge that can't be expressed by means of words and so it is difficult to be formalized and communicated.

The outspoken knowledge, instead, can be captured and codified in manuals, in procedures or in rules; it can be therefore delivered by means of a formal and systematic language.

Most of the knowledge management effort consists in turning the unspoken knowledge in new corporate assets that can be conveyed to all its components. Moreover, knowledge and information are not produced inside a single organization, but a number of different communities can produce them. Information and ideas are shared and within an organization, becoming knowledge that will be shared by the whole community which the organization belongs to.

Therefore, the knowledge is distributed between various organizations that belong to a community and it can be very heterogeneous; in particular, it could be different, more careful and potentially contradictory if we compare the same one form different organizations' points of view.

The knowledge management philosophy is put in action by sharing this knowledge among different groups inside an organization or, more in general, between various organizations that belong to the same community.

The process of knowledge sharing allows creating new knowledge that, otherwise, will not be created. This cross-fertilization mechanism is not a mere contamination directed to the same one, but it is a creative sharing process, since the interaction inside a community allows the creation of new knowledge and innovative ideas. These generic considerations find a perfect applicability in the area of cultural assets, especially in Italy [1], where the knowledge belongs to various subjects, public and private (the Church, the Ministries, private collectors, researchers, etc.). They are scattered on all of the national territory and often done not coordinate. The current scenery reveals an enormous proliferation of information sources that are local, scattered, heterogeneous, little known, barely accessible, scarcely reused, and hardly reusable. If, today, to professional wants to collect all the information about to possible item of interest (say to specific piece of ceramics or an archaeological finding), s/he must look at different sources: public given bases, catalogues of exhibitions, art books, scientific papers, etc.

The sharing and the integration of information between all members of a community happen by means of exchange processes that can be formal or informal. For example, a member of a community who is looking for any type of information could contact directly other groups which potentially own information (broadcasting), or it could identify an element that interpose itself between all members and which function is to catalyze and organize the distributed knowledge (the broker). In the last case, only the broker knows who owns certain knowledge; it further shares, with all members of the community, the meaning of the words and it knows how to map all concepts with each one's knowledge. What we want to realize, and that partially has been already realized in the DICE project (Distributed Infrastructure for Cultural hEritage), is a technological and organizational system which allows members providing a virtual space where the knowledge comes and is shared from all members of the community. In DICE, the sharing and the integration is not a mere sharing of information that belongs to every source, but the system supplies a structure that allows organizing the knowledge. DICE supplies an ontology of this knowledge base that allows each member of a community to share, to represent, to organize, to research and to associate their own information with other members' information.

Furthermore, from the organizational point of view, taking into account the past failures [2,3,4], DICE implements a bottom-up integration model. In the DICE vision, a number of DICE communities will rise up around specific topic. In each community, users agree on a particular ontology (called "cultural model" in the following) and, in the future, different communities will be integrated by means of other ad-hoc cultural models.

Strategic goal of the DICE project is to encourage a sustainable valorisation of the Italian cultural heritage, through the development of virtual communities. Communities will integrate different professional user profiles (e.g., researchers and scientists, promoters of tourism, editors, etc.), who will participate by sharing their own information and acting as "cultural mediators" towards the final users.

On the other hand, 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.

In particular, the infrastructure design has been driven by a set of requirements, assumptions and constraints that can be summarized in the following list:

1. ***Information ownership***: each participant of a DICE community must keep the ownership of its information. This is a fundamental and psychological requirement because information is the most valuable resource of each participant.
2. ***Joint community***: within a DICE community all members have the same power, that is, the community survival cannot be subordinated to the existence of a particular member. In a community there is not a single broker but all members share the knowledge and it do not go lost in case any participant decides to leave the community.
3. ***Cultural model agreement***: when a new information source joins a DICE community it must agree to the relative cultural model.
4. ***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.
5. ***Scalability***: The infrastructure should be easily scalable in terms of number of participants of a community.

The paper discusses in details the conceptual and practical problems detected in two field studies that led to the creation of a DICE "demonstrator" implemented in March 2004. More than 20 different sources both for "Archaeology in Campania" and for "Ceramic in Campania" were integrated, 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. A "user centred" approach to the integration of such sources is presented, in that the integration strategy has been tailored to the user needs rather than to the data characteristics (like in most of existing approaches). Based on what the user is willing to do, and what (s)he is available to do, we have defined an integration approach and a technological infrastructure based on probabilistic algorithms exploiting the shared knowledge in a cultural community.

In the following paragraph we briefly describe the DICE Cultural Model that will be deepening in the paper.

2. The Cultural Model

The cultural model 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?”

All concepts that belong to the "cultural model" and that we will describe in the following are: information source, content and profile schema, taxonomy, information unit, semantic association, access path, guided access. Moreover, as a distinctive characteristic of the DICE approach, “intelligent” data extraction algorithms and link setting probabilistic algorithms are defined in DICE to set up the integration knowledge base.

Information sources (IS) could be public databases, catalogues of exhibitions, art history books, scientific papers, etc. In particular, we assume that:

- Different sources may use different formats (both in terms of content and in terms of database structure) to represent their own information
- Different sources may use different languages (different terminology) to describe their information
- Different sources (even if they use the same metadata thesaurus) may have different ways of characterizing objects (either by mistake or by representing different opinions)
- Different sources may have different ways to identify objects (unofficial sources do not use official identifiers; many objects do not have official identifiers anyway)

In order to define the visual structure of information about cultural assets we introduce the concept of *content schema* (CS). Its purpose is to structuring information in terms of paragraphs of content and it is designed in order to provide a uniform visualization for similar information. Within the same DICE community, a range of different content schemas are defined and when a new information source joins to a community a content schema has to be chosen for representing its own information. It is not unusual that two information sources use different paragraphs to describe the same typology of information.

While the content schema is introduced to represent how to visualize information about cultural assets, the *profile schema* (PS) is introduced to characterize them.

A profile schema consists of a set of categories and all possible value that can be associated to each category of the profile schema, are taken by a corresponding *taxonomy* (a structured vocabulary of terms connected by a specialization relationship).

When a new information source joins to a community, a profile schema has to be defined for characterizing its information.

It should be noticed that while the purpose of the content schema is to structuring information for user consumption, the profile schema is only used for setting links among different cultural assets (that share some common characteristics) and to allow the user to find them.

Related to information sources, with their content and profile schemas, the concept of information unit (IU) must be introduced to represent all possible pieces of information that they share. An IU could be an index card of a museum, a record of database, a picture of a photo library and it is considered in DICE the atomic unit of information available for the user. Consequently, it follows that an IU belongs to an IS, it is made of a content, instance of the IS’s CS, and a profile, instance of the IS’s PS.

In particular, it can be observed that separation between profile and content of an IU allows having two levels of integration, the visual level and the content level. At the visual level, a uniform representation of contents allows users to improve their “user experience”, which is improving the application usability and acceptability. At the content level, profiles allow the integration engine to find information out of a user need.

Once information has been represented, a number of access mechanisms should be defined in orded to allow users reaching interesting information units. Considering any interactive application exploiting the navigational interaction paradigm, the access mechanisms can be divided into three main categories:

- **Predefined navigational paths:** it provides users with predefined and guided ways to discover interesting objects on the basis of supposed user interests or needs. They are typically used at the beginning of a navigation session corresponding to some user task.
- **Navigational semantic associations:** once a user has landed on an interesting object, this access mechanism allows (her)him navigating towards other related and potentially interesting objects.
- **Search engine:** when users are unable to find out interesting objects by means of previous access mechanisms or when they have a clear and specific need, it provides a versatile means for accessing the overall information base. On the other hand, due to its generality, such a mechanism is ineffective in proposing new key to the reading and is usually unsuitable for naive users.

Our approach embodies all these access mechanisms specially suited for the specific problem and domain. In particular, concerning *predefined navigational paths,* in our approach in each community a panel of cultural experts defines a set of them. Defining them, a number of factors like cultural topics, kind of users, and so forth, are considered.

Concerning *navigational semantic association,* so far we have identified three different typologies of associations:

- **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).

It is important to notice that intelligent algorithms generate all links (corresponding to semantic association instances). Founded on profiles of IUs and on taxonomies, they are specific for each semantic association and they assign a value (from 0 to 1) to each link; this value represents the probability that the specific semantic association relates two IUs.

At the end, we can summarize that the ultimate purpose of the cultural model is to transform each member's implicit knowledge in explicit knowledge that is shared by all members of a community.

In particular, content schema introduces new knowledge that is related to the structuring of information in terms of paragraphs.

Profile schema, on the other hand, allows translating implicit knowledge related to content in explicit knowledge by means of categories and terms.

Taxonomies make the implicit base knowledge of the community explicit. Taxonomies define all possible terms and relation between them that community members usually use to refer to information.

Finally, predefined navigational paths and navigational semantic associations make relations between information explicit. In particular, navigational paths can be defined by each participant that decides to share their knowledge about relations between information units (they belong to the same navigational path) with all others members of the community.

On the other hand, in the case of navigational semantic associations the knowledge about relations is implicit in the semantic of algorithms that allow to setting links.

References

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

[2] 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

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

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