Digital libraries of the future – and the role of libraries

Donatella Castelli
ISTI-CNR, Pisa, Italy

Abstract

Purpose: To introduce the digital libraries of the future, their enabling technologies and their organizational models.

Design/methodology/approach: The paper first discusses the requirements for the digital libraries of the future, then presents the DILIGENT infrastructure as a technological response to these requirements and, finally, it discusses the role that libraries can play in the organisational framework envisioned by DILIGENT.

Findings: Digital libraries of the future will give access to a large variety of multimedia and multi-type documents created by integrating content from many different heterogeneous sources that range from repositories of text, images, and audio-video, to scientific data archives, and databases. The digital library will provide a seamless environment where the co-operative access, filtering, manipulation, generation, and preservation of these documents will be supported as a continuous cycle. Users of the library will be both consumers and producers of information, either by themselves or in collaborations with other users. Policy ensuring mechanisms will guarantee that the information produced is visible only to those who have the appropriate rights to access it. The realisation of these new digital libraries requires both the provision of a new technology and a change in the role played by the libraries in the information access-production cycle.

Practical implications: Digital libraries of the future will be core instruments for serving a large class of applications, especially in the research field.

Originality/value: The paper briefly introduces one of the most innovative technologies for digital libraries, and it discusses how it contributes to the realization of a novel digital libraries scenario.

Keywords: Digital libraries, Knowledge management

Paper type: Viewpoint

Introduction

Research on digital library (DL) systems started in Europe in the mid-nineties. At that time DLs were seen essentially as *repositories of digital texts* accessible through a *search service* which was operating by indexing information stored in a *centralised metadata catalogue*. The construction of a DL was very resource-consuming since, for each new DL, both the content and the software providing the DL functionality were built from scratch. As a result of this development approach, only powerful user communities [1] or user communities with in-house computer science technical skills (Leiner, 1998) could afford the building up of DLs. These DLs were created to serve end-users only as *consumers* of information. They did not provide any functionality for submitting the documents. The submission was usually performed either by the author or by a librarian operator by means of specific procedures residing outside the DL.

Today, the requirements imposed on DLs are very different from that early time. A novel notion of DLs, also referred to as “knowledge commons” (Ioannidis, 2005), has recently emerged, whose fulfilment requires new technologies and new organisational models. This paper focuses on such new DLs by first discussing the motivations for their introduction, then presenting an innovative DL technology, called DILIGENT, and, finally, illustrating the role that libraries can play in this new scenario.

Digital libraries of the future

According to the most recent understanding, the DLs of the future will be able to operate over a large variety of information object types - far wider than those maintained today in physical libraries and archives. These information objects will be composed of several multi-type and multimedia components aggregated in an unlimited number of formats. These, for example, can mix text, tables of scientific data and images obtained by processing earth observation data, or they can integrate 3D images, annotations and videos. These new information objects will offer innovative and more powerful means to researchers for sharing and discussing the results of their work. In order to be able to support these objects, the DL functionality has to be appropriately extended far beyond that required to manipulate the simple digital surrogates of the physical objects. In order to support these objects the DL may need considerable resources. For example, the creation and handling of the new documents may require access to many different, large, heterogeneous information sources, the use of specialised services that process the objects stored in these sources for producing new information, and the exploitation of large processing capabilities for performing this tasks.

New DLs are also required to offer a much richer set of services to their users than in the past. In particular, they must support the activities of their users by providing functionalities that may range from general utilities, like annotation, summarization or co-operative work support, to very audience-specific functions, like map processing, semantic analysis of images, or simulation. The availability of this new DL functionality can, in principle, change the way in which research is conducted. By exploiting such types of DL, for example, a scientist can annotate the article of a colleague with a programme that extracts useful information from a large amount of data collected by a specific scientific observatory. This programme, executed on demand when the annotation is accessed, can complement the content of the paper with continuously refreshed information.

In the new DLs users are not only consumers but also *producers* of information. By elaborating information gathered through the DL they can create new information objects that are published in the DL, thus enriching its content. The new DLs are thus required to offer services that support the authoring of these new objects and the workflows that lead to their publication.

In parallel with the above evolution of the role of DL systems, we are now observing a large expansion in the demand for DLs. Research today is often a collaborative effort carried out by groups belonging to different organizations spread worldwide. Motivated by a common goal and funding opportunities, these groups dynamically aggregate into virtual research organizations that share their resources, e.g. knowledge, experimentation results, or instruments, for the duration of their collaboration, creating new and more powerful virtual research environments. These virtual research organizations, set up by individuals that do not necessarily have great economic power or technical expertise, more and more frequently require DLs as tools for accelerating the achievement of their research results. This new potential audience demands less expensive and more dynamic DL development models. They want to be able to set up new DLs that serve their needs for the duration of their collaborations in an acceptable timeframe and with an acceptable cost. The current DL development model cannot satisfy this large demand; a radical change is needed if we want to be able to address these new emerging requirements.

A great contribution towards the satisfaction of all the above mentioned requirements can certainly come from the introduction of mechanisms that support a *controlled sharing of resources* among different organisations. Sharing in this context is not only applied to repositories of content, as is usually meant today, but can be extended to any type of resource needed to build a DL, i.e. language and ontology resources, applications, computers and even staff with the necessary skills for supporting the DL development, deployment and maintenance. Supporting this type of sharing requires the introduction of appropriate solutions at both the *technological* and *organisational* levels. These two levels are not independent; instead they strongly influence each other. In fact, the availability of a good technological solution favours the creation of an appropriate organization, and vice-versa, a successful organization stimulates the development of new supporting technologies.

In the next section we present the DILIGENT infrastructure as an example of a technological solution for these new DLs. The organisational aspects stimulated by the introduction of this technology are briefly discussed afterwards.

DILIGENT

DILIGENT (DIgital Library Infrastructure on Grid ENabled Technology) [2] is a three-year Integrated Project (2004-2007) funded by the European Commission under the 6th Framework Programme for Research and Technological Development. The objective of this project is to develop a Digital Library Infrastructure that will enable members of dynamic virtual research organizations to create on-demand transient digital libraries that exploit shared resources. Resources in this context are multimedia and multi-type content repositories, applications, and computing and storage elements. Following the understanding of DLs expressed in Borgman *et al.* (2002), this project focuses on the development of DLs that "are not ends in themselves; rather they are enabling technologies for digital asset management, electronic commerce, electronic publishing, teaching and learning, and other activities" (p. 7).

From an abstract point of view, the DILIGENT infrastructure can be understood as a broker serving DL resource providers and consumers. The providers are the individuals and the organizations that decide to publish their resources under the supervision of the broker, according to certain access and use policies. The consumers are the user communities that want to build their own DLs. The resources managed by this broker are content sources (i.e. repositories of information searchable and accessible through a single "entrance"), services (i.e. software tools that implement a specific functionality and whose descriptions, interfaces and bindings are defined and publicly available) and hosting nodes (i.e. networked entities that offer computing and storage capabilities and supply an environment for hosting content sources and services). Providers register their resources and give a description of them by exploiting appropriate mechanisms provided by the infrastructure. The infrastructure also automatically derives other properties of the resources that are used to enrich the explicit description. The infrastructure manages the registered resources by supporting their discovery, monitoring and usage, and by implementing a number of other functionalities that aim at realising the required controlled sharing and quality of service. A user community can create one or more DLs by specifying a set of requirements. These requirements specify conditions for the information space (e.g. publishing institutions, subject of the content, document types), for the operations that manipulate the information space (e.g. type of search, tool for data analysis), for the services for supporting the work of the users (e.g. type of personalized dissemination, type of collaboration), for the quality of service (e.g. configuration, availability, response time) and for many other aspects, like the maximum cost, or lifetime. The broker satisfies the community's requirements by selecting, and in many cases also deploying, a number of resources among those accessible to the community, gluing them appropriately and, finally, making the new DL application accessible through a portal. The composition of a DL is dynamic since the DL broker continuously monitors the status of the DL resources and, if necessary, changes them in order to offer the best quality of service. By relying on the shared resources many DLs, serving different communities, can be created and modified on-the-fly, without big investments and changes in the organizations that set them up.

In order to support the transactions between the providers and the consumers, the DILIGENT infrastructure exploits the virtual organizations (VOs) mechanism that has been introduced in the Grid research area (Foster *et al.*, 2001). This mechanism models sets of users and resources aggregated together by highly controlled sharing rules, usually based on an authentication framework. VOs have a limited lifetime, are dynamically created and satisfy specific needs by allocating and providing resources on demand. Through the VOs mechanism the DILIGENT infrastructure glues together the users and the resources of a DL.

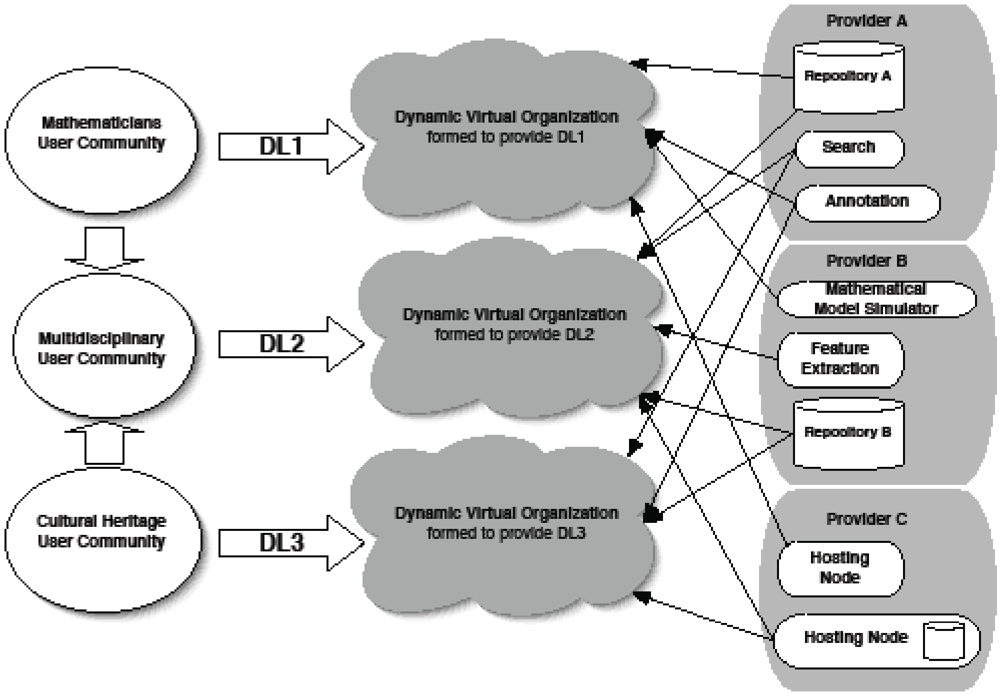
Mathematicians User Community
Multidisciplinary User Community
Cultural Heritage User Community
DL1
DL2
DL3
Dynamic Virtual Organization formed to provide DL1
Dynamic Virtual Organization formed to provide DL2
Dynamic Virtual Organization formed to provide DL3
Provider A
Repository A
Search
Annotation
Provider B
Mathematical Model Simulator
Feature Extraction
Repository B
Provider C
Hosting Node
Hosting Node

Figure 1: The role of virtual organizations in DILIGENT

Figure 1 graphically illustrates the role of VOs in supporting the brokerage model. The consumers, i.e. the user communities that require DLs to support their needs, are on the left of the figure. The providers, each of which makes a number of resources available, are on the right. The infrastructure acts as a mediator by maintaining a framework where multiple virtual organizations, active on the same shared resources, can co-exist.

The DL development model proposed by DILIGENT is radically new. Within the described framework each DL consumes the required resources only for the time it needs them. This opens a lot of new opportunities for the creation of the functionalities required by the new “knowledge commons” environments. In particular, the exploitation of more effective, but also very computationally expensive algorithms becomes viable at an acceptable cost for many communities. For example, thanks to sharing, the use of the high process-consuming algorithms that automatically extract features from multi-media objects can be exploited in a large number of DLs. Moreover, in the framework established by the new development model, the user communities can easily, and in a timely manner, create and maintain their own DLs with limited resources since the management of the DL is automatically and transparently carried out by the infrastructure.

The system that implements the functionality of the DILIGENT infrastructure is being built by integrating DL and Grid technologies (Foster and Kesselman, 2004). The motivation for this design choice relies on the similarity between many of the problems encountered through our new notion of DLs and the issues addressed by the most recent research in the Grid domain.

From the functional point of view, the DILIGENT system is divided into five functionality clusters:

1. *DL Creation & Management:* is responsible for the dynamic construction and maintenance of the transient DLs and for the controlled sharing and management of the resources that are used to implement them. The functionalities offered by this cluster allow users to express the requirements that the DL must fulfil. Moreover, they automatically identify and arrange the pool of resources needed to satisfy these needs.
2. *Content & Metadata Management:* implements the handling of DL content and related metadata, the consistent and distributed management of annotations, and the integration of external content and metadata sources.
3. *Process Management:* manages the creation of user processes composed of existing services, the validation of their correctness, the automatic optimisation of their definition according to the resources available and the service characteristics, and their reliable execution. Thanks to this feature, the DILIGENT system can easily be enriched with additional operational workflows to meet new user requirements.
4. *Index & Search Management:* is responsible for enabling cost-efficient search and retrieval of information in DLs, while satisfying the level of quality required for the overall data retrieval and delivery operations.
5. *Application Specific Functionality:* provides the functionality needed to support user-specific scenarios, like portals, document visualization, or features extraction.

From the architectural point of view, the DILIGENT system is designed as a Web Services Resource Framework (WSRF) application (Foster *et al.*, 2004) built on top of the gLite Grid middleware [3] released by the Enabling Grids for E-science in Europe (EGEE) project [4]. gLite hides the heterogeneous nature of the computing elements (i.e. services representing a computing resource) on the one hand and storage elements (i.e. services representing a storage resource) on the other hand by providing an environment that facilitates and controls their sharing.

The DILIGENT services are being initially deployed on a project-proprietary gLite infrastructure. Architecturally, this infrastructure is completely interoperable with the EGEE infrastructure. EGEE is currently the largest European Grid infrastructure ever built. A number of other recently funded projects will extend this infrastructure to other geographic regions, like Mediterranean countries, Latin America, or China. The interoperability with the EGEE infrastructure will allow any authorised virtual community that wishes to create DLs to also exploit the resources made available by this vast Grid infrastructure.

During the project timeframe, the DILIGENT infrastructure will be populated with a number of important archives and software applications provided by the two communities that are participating in the experimentation with the results of the project, one from the environmental e-science domain and one from the cultural heritage domain. The first community is ImpECt (Implementation of Environmental Conventions) and includes leading players in the environmental sector. This community will use DILIGENT to support the organization of conferences and the preparation of projects and periodical reports. Through DILIGENT this community expects to improve accessibility, interoperability and usability of environmental data, models, tools, algorithms and instruments, integrating the distributed data sources with specialized data handling services. The second community, ARTE, is a community of scholars located in different parts of the world, working together to establish a new discipline that merges experiences from research in medicine, humanities, social sciences and communication. In order to achieve their objectives, these researchers require instruments to ease the construction of multimedia artefacts and to improve support for education.

At the time of the writing of this paper (February 2006), initial experimentation with the features of a DILIGENT DL has already been conducted by implementing simplified services for preparing environmental reports, as required by the ImpECt environmental agencies. Through the exploiting of rich information sources, ranging from raw data sets to maps and graphs archives, these agencies periodically prepare reports on the status of the environment. Currently, this task is performed by first selecting the relevant information from each of the multiple and heterogeneous sources available, then launching complex processing on large amounts of data to obtain “products”, like graphs, tables and other summarized information and, finally, producing the required report by assembling all the different parts together. This process, which is repeated periodically, requires a lot of work due to the complexity of interfacing the different sources and tools. Despite the effort expended, the resulting reports do not completely fulfil the requirements of their readers, who would like to have a picture of the environmental status which is updated at the time the report is accessed. A DILIGENT DL offers a more effective framework for the creation and maintenance of these reports. In our experimentation, for example, we have built a DL which exploits content maintained in both repositories including textual documents and archives of Earth observation raw data provided by the European

Space Agency. In this DL, which is accessible through a single user interface, all the different kinds of information necessary for creating the reports can be found. By combining this information and by defining how to derive the associated "products" (images, tables, or graphs) from raw data, the users can create their reports much more easily. Moreover, a specialized user interface allows authorised users to access these composite reports by choosing static or dynamic generation. The selection of dynamic generation triggers associated process workflows which, by combining appropriate applications, generate the required products on demand by processing both the raw data and other intermediate products stored in the DL repositories.

The dynamically generated products are obtained by running those applications that are computationally intensive on the Grid. In this way the complex processes required to generate the products are executed in few minutes at a limited cost to the community that is exploiting them. In order to obtain the same performance without the Grid, an institution would have to equip its digital library with a great number of computers, while in the case of DILIGENT the institution can also exploit computer capabilities made available by third party organisations. The same is true for storage capacity. Maintaining raw data, intermediate products and high resolution images requires a large amount of storage capacity. By exploiting the Grid technology, part of this information, especially the temporary part, can be maintained in third-party storage systems.

The role of libraries in future DLs

In the framework envisaged by DILIGENT, libraries play an important role at the organizational level. In particular:

- As providers of resources, they can help to enhance the amount of available resources by making stakeholders aware of the importance of sharing. In particular, as far as the sharing of content is concerned, they can operate by promoting digitisation campaigns and the Open Access approach. These actions may result in a vast amount of new digital information accessible online which can be exploited by advanced services.
- Also within a digital framework, libraries are certainly the best candidates for carrying out content description, maintenance and preservation of resources. By exploiting their large experience acquired in the past, they can contribute to the long-term availability and to the quality of the resources disseminated by the DLs.
- Long-term availability also requires the implementation of models able to support the sustainability of the resources provided. Libraries, either alone or as members of library consortia, can also act as the organisations deputed to define and put in place these models.
- As main resource providers, libraries can work jointly on the definition of common policies and standards. An agreement on these aspects would strongly contribute towards facilitating the design and development of the new complex services required to fulfil the emerging user needs.
- In the future envisaged by DILIGENT, libraries can also play an important role as mediators between the infrastructure and the user communities. In particular, they can proactively promote and facilitate the creation of DLs that respond to the needs of the user communities. They can also assist

users by providing, if necessary, the skills required to select, update and exploit the DL content and services.

Concluding remarks

This paper has introduced a vision of the DLs of the future and it has presented DILIGENT, a new technology which is being developed to support this new vision. It has also discussed the role which libraries can play in the framework envisaged by DILIGENT by outlining, especially, their contribution at the organisational level.

At the time of the writing of this paper, the DILIGENT project is halfway towards the achievement of its objectives. As outlined in the paper, few concrete experiments have already been done to test new functionality and the impact that this new technology may have on supporting the work of specific user communities. The reaction of the user communities involved in these experiments is very encouraging. Other communities, especially scientific communities, have asked to test the technology developed since they found that the DL model proposed can dramatically change the way in which their activities are conducted.

As outlined at the beginning of this paper, the technology by itself is not sufficient to implement the new model envisioned by DILIGENT. A consistent effort is also needed at the organisational level. It is on this level that libraries can play a key role by bringing their valuable experience to bear in the new scenario.

Notes

1. For example the ACM Digital Library: http://portal.acm.org/dl.cfm
2. www.diligentproject.org
3. http://glite.web.cern.ch/glite/
4. http://public.eu-egee.org

References

Borgman, C., Sølvberg, I. and Kovács, L. (Eds.) (2002), *Proceedings of the Fourth DELOS Workshop, Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics*, Budapest, 6-7 June 2002, available at: www.sztaki.hu/conferences/deval/presentations/DELOSWorkshop4OnEval_report.pdf (accessed 4 May 2006).

Foster, I. and Kesselman, C. (Eds.) (2004), *The Grid: Blueprint for a Future Computing Infrastructure*, 2nd ed., Kaufmann, Amsterdam.

Foster, I., Kesselman, C. and Tuecke, S. (2001), "The anatomy of the Grid: enabling scalable virtual organizations", *The International Journal of High Performance Computing Applications*, Vol. 15, No. 3, pp. 200–222.

Foster, I., Frey, J., Graham, S. and Tuecke, S. (Eds.) (2004) *Modeling Stateful Resources with Web Services*, Version 1.1, 3 May 2004, Whitepaper, available at: www-128.ibm.com/developerworks/library/ws-resource/ws-modelingresources.pdf (accessed 4 May 2006).

Ioannidis, Y. (2005), "Digital libraries from the perspective of the DELOS Network of Excellence", in *Proceedings of the IEEE-CS International Symposium:*

Global Data Interoperability - Challenges and Technologies, June 20th - 24th, 2005, Sardinia, Italy, pp. 51-55, available at: http://globalstor.org/ (accessed 4 May 2006).

Leiner, B.M. (1998), "The NCSTRL approach to open architecture for the Confederated Digital Library", in *D-Lib Magazine*, Vol. 4, No. 11, available at: www.dlib.org/dlib/december98/leiner/12leiner.html (accessed 4 May 2006).