A Semantic Web Powered Distributed Digital Library System

Michele Barbera[1]; Michele Nucci[2]; Daniel Hahn[1], Christian Morbidoni[2];

[1] Net7 – Internet Open Solutions -
Via Marche 8/a, 56123 Pisa, Itay
e-mail: barbera@netseven.it; hahn@netseven.it
[2]Dipartimento di Elettronica, Intelligenza artificiale e Telecomunicazioni, Università politecnica delle Marche,
Via Brecce bianche, 60100 Ancona, Italy
e-mail: mik.nucci@gmail.com; christian@deit.univpm.it

Abstract

Research in Humanities and Social Sciences is traditionally based on printed publications such as manuscripts, personal correspondence, first editions and other types of documents which are often difficult to obtain. An important step to facilitate humanities and social sciences scholarship is to make digital reproductions of these materials freely available on-line. The collection of resources available on-line is continuously expanding. It is now required to develop tools to access these resources in an intelligent way and search them as if they were part of a unique information space. In this paper we present Talia, a innovative distributed semantic digital library, annotation and publishing system, which is specifically designed for the needs of scholarly research in humanities. Talia is strictly based on standard Semantic Web technologies and uses ontologies for the organization of knowledge, which can help the definition of a highly adaptable and state-of-the-art research and publishing environment. Talia provides an innovative and flexible system which enables data interoperability and new paradigms for information enrichment, data-retrieval and navigation. Moreover, digital libraries powered by Talia can be joined into a federation, to create a distributed peer-to-peer network of interconnected libraries. In the first three paragraphs we will introduce the motivations and the background that led to the development of Talia. In paragraphs 4 and 5 we will describe the Talia's architecture and the Talia Federation. In paragraphs 6 and 7 we will focus on Talia's specialized features for the Humanities Domain and its relations with the Discovery Project.In paragraph 9 we will describe Talia's widget framework and how it can be used to customize Talia for other domains. In the final paragraph we will compare Talia with related technologies and platforms and suggest some possible future research and development ideas.

Keywords: digital library; semantic web; humanities.

1. Introduction

In the last few years the amount of digital scholarly resources in the Humanities grew substantially thanks to the efforts of many collections holders who digitized their materials and published them on-line. However, to date, many digital library projects can be characterized as both strongly hierarchical (top-down) and disconnected. Materials are selected for reformatting and inclusion by librarians, archivists and curators from their own collections to the presumed benefit of their patrons but with little actual consent from them. Collections are often assembled with little regard for existing complementary materials, leaving it to the end-user to make and sustain the connections across collections, that remain collections remain fundamentally siloed, with no way to establish permanent semantic connections of their contents. In digital research libraries there is no longer any need to abide by the restrictions of physical organizational schemes or even physical location. New research libraries can and should be built across collections and across libraries. On the other side of this spectrum lay the large-scale aggregation initiatives, such as the The European

Library [1] or OAIster [2] which serve as general purpose digital libraries, but fail to provide the depth needed for research-level scholarship.

The emergence of web 2.0 has resulted in a number of tools and technologies for annotation and personalization of resources but these tools have yet to gain a strong foothold in an humanistic academic setting. We believe that Semantic Web Technologies have the potential to glue together the opposing needs of maintaining the context in which the collections originate, by leaving them under control of their holders, and at the same time making resources part of a global structured knowledge space that is independent of a single centralized authority or aggregation service.

2. Semantic Web and Ontologies

The Semantic Web is an extension of the current Web in which information can be expressed in a machine-understandable format and can be processed automatically by software agents. The Semantic Web enables data interoperability, allowing data to be shared and reused across heterogeneous applications and communities [3].

The Semantic Web is mainly based on the Resource Description Framework (RDF) [4] by which is possible to define relations among different data, creating semantic networks. RDF's main strength is simplicity: it is a network of nodes connected by directed and labelled arcs (figure 1).

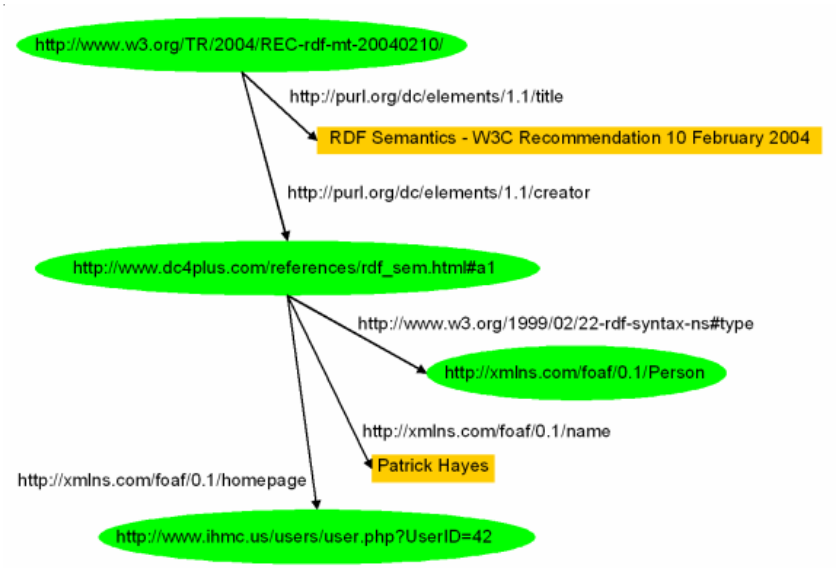


Figure 1: An example of an RDF Semantic Network

The nodes are resources or literals (values) while arcs are used to express properties of resources. In SW a resource is anything that can be somehow identified using a specific identifier. In particular, the identifiers used in SW are known as Uniform Resource Identifiers (URI).

In Semantic Web ontologies, are used to organize information and formally describe concepts of a domain of interest. An ontology is essentially a vocabulary which includes a set of terms and relations among them. Ontologies can be developed using specific ontology languages such as: the RDF Schema Language[1] (RDFS) and the Web Ontology Language[2] (OWL).

3. Digital libraries for all

In recent years, the decreasing prices of digitazion costs and storage facilities as well as the emergence of many easy-to-deploy Open Source content management systems and digital object repositories, led to the multiplication of small digital libraries run by smaller institutions. Despite their limited size, the collection of these libraries sometimes include cultural masterpiecies. Unfortunately, due to limited resources, these libraries cannot afford to invest on professional digital library management platforms that are either too expensive in terms of license costs or too expensive to maintain because of their complexity. Talia is an Open Source semantic digital library management system that is easy to deploy and maintain. Building a Talia based digital library doesn't require any advanced software development and management skill, that smaller cultural institutions may not possess. Additionally, Talia is a distributed library system, meaning that it permits to build virtual collections that go beyond the boundaries of a single archive without requiring the underlying content providers to loose control over their holdings. For all the reasons stated above, Talia aims at being a complete tough powerful solution for the needs of smaller institution's digital libraries.

4. The Talia Platform

Talia is a distributed semantic digital library system which has been specifically designed for the needs of scholarly research in social science and Humanities. Talia combines the features of a digital archive management system with an on-line peer review system, thus it is capable of combining together a digital library with an electronic publishing system. Talia is able to handle a wide range of different kinds of resources such as texts, images and videos. All the resources published in Talia are identified by a stable URI: documents can never be removed once they are published and are maintained in a fixed state in perpetuity. This, combined with other long-term preservation techniques, allows the scholars to reference their works and gives the research community immediate access to new contents.

One of the most innovative aspect of Talia is that it is completely based on Semantic Web Technologies which enable deep data interoperability with other Semantic-Web aware tools and applications. In particular, the Talia Knowledge Base is kept in RDF and it is formally described using RDFS/OWL ontologies. Talia natively supports heterogeneous data sets whose metadata schemes can be very different from each other, therefore the system is not based on a predefined ontology. Talia embeds only a very broad structural ontology, which contains only general concepts and basic relations to link resources. Research communities are encouraged to develop their own domain ontology to describe knowledge and content in their domains of interest. The domain ontologies can be developed using standard ontology languages such as RDFS and easily imported into into the library's data store.

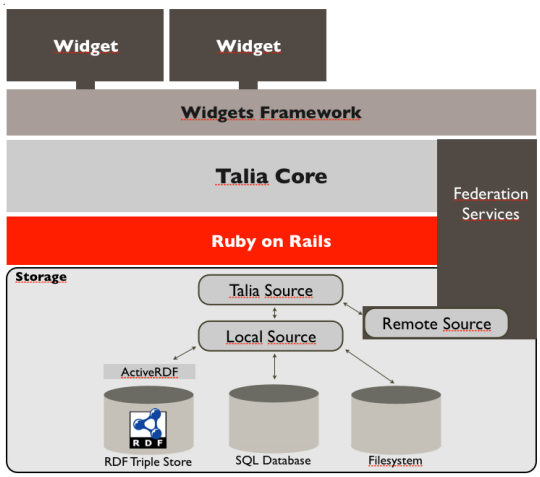


Figure 2: Talia Architecture

Talia also includes facilities for semantic data-enrichment, data-annotations and data-retrieval as well as a lot of other specific tools. A complete overview of the Talia architecture as well as the underlying technical details can be seen in [5] and [6].

5. Distributed semantic digital libraries

Digital semantic libraries based on Talia can be joined in a Talia Federation, to create a peer-to-peer network of interconnected libraries (nodes). Talia provides a mechanism to share parts of its knowledge base. This mechanism is based on a REST interface, an approach proposed in [7]. By using this feature a node can notify another node that a semantic link between them exists. This information can then be used by the notified node to update its own knowledge base to create a bidirectional connection between the contents. This features allows individual scholarly communities, each one managing a single node of the federation, to retain control on their own content while at the same having a strong interaction with the other nodes content. Talia works both as a Digital Library and as an Open Access publisher of original contributions. In a digital library a notion of absolute quality can be acceptable even outside the boundaries of the community who manages the library. On the other hand, the concept of quality for newly published contributions varies significantly with culture and context, therefore each communiy must retain control on what their users see through that community web site. The approach used in Talia is to let each node decide which other federation node it trusts. Notification of incoming links will then be processed only if they come from trusted nodes. The result for the enduser is that it sees backlinks only to content held in trusted sources. At any time, a node administrator can modify the trust policy and recover the notifications that have been previously filtered out by the trust engine.

Talia also features a single sign-on mechanism based on OpenID [8]. A Talia node acts as an OpenID client. Depending on its own policies a federation can run it's own OpenID identity server or choose to rely on any existing external service. Each federation node keeps a copy of the user credentials and user roles and permissions are managed locally. As any other of its components, the authentication and authorization component of Talia is pluggable and completely modular. Therefore, it will be possible in the future to develop specific authentication and authorization components based on other infrastructures, such as a more institution-centric approach based on the Shibboleth [9] model, or on any other legacy model.

6. Digital Humanities

Projects in the domain of Digital Humanities deal with a incredible amount of different types of data (ranging from manuscript reproductions to statistical linguistic metadata, pictures of historically relevant places, maps and many different kind of books in diverse digital formats just to name a few). The level of standardization of data and metadata formats, but especially of the research process is very limited compared with natural sciences. Another important characteristic of this domain, is that the level of computer literacy for Humanities scholars tends to be rather low compared to scholars in other sectors. These two facts (heterogeneity of data and processes and low computer literacy of the users) suggest the need of an electronic environment that must be extremely flexible, adaptable and extensible but at the same time integrated and easy to use.

Talia is a coherent and easy-to-use web-based working environment that integrates a set of features that are usually scattered through many different desktop and web-based tools rather than condensed in a unique environment. These tools include for example XML tagging and transformation, image and audio analysis, linguistic text analysis, manuscript annotation, electronic edition editing and so on.

Talia's widget system permits to extends the core engine and easily integrate these tools into a unique

infrastructure. In the context of the Discovery project we will develop a limited amount of these tools as well as the documentation on how to develop additional plug-ins. We hope that the Open Source community and other Digital Library projects will contribute additional plug-ins in the near future.

7. The discovery project

The Discovery project [10], funded by the European Commission under the eContentplus programme, aims at the creation of a federation of digital libraries dedicated to different authors and themes of ancient, modern and contemporary philosophy.

Talia has born in the context of the Discovery project to serve as its technological infrastructure. The federated libraries that are part of the Discovery federation are unique as they as the function at the same time as traditional digital libraries and as Open Access publishers of original contributions. With this model , discovery aims at stimulating the production of new knowledge by aggregating scholarly communities around thematic repositories of both primary sources (like manuscripts and first editions) and original contributions submitted by the scholars. Additionally, the nodes of the Discovery federation, can also store and publish semantic annotations, that are another type of user contributions. In Discovery, there are two main categories of resources (called "Sources" in Discovery): primary and secondary sources. Primary sources are all the resources that belong to the digital library. These resources have been collected, digitized and published by the institution that runs the digital library. Secondary sources are all the resources that belong to the Electronic Publisher component of a Talia node. These resources have been submitted by the users and they passed through a peer-review process before being published.

In Discovery there are four content providers, each of them manages is own instance of Talia. Each provider has its own Domain Ontology that specifies wich types of Primary and Secondary Sources they deal with. Each content provider also has its own peer-review policies and procedures and user user interfaces. In addition to running domain specific Digital Libraries and an Open Access Electronic Publishers, the content providers also engages in what is refferred to as "Semantic Enrichment".

By using a tool called Philospace[10], domain specialists can semantically annotate the Sources published in Talia to add new semantic relations among them. As any other user generated content, the semantic annotations also go through the peer review process. If the annotations pass the peer-review scrutiny they are published into Talia and become available to end users. There is no limit on the meaning of the annotation that may range from simple metadata added to a Source to complex relations to philosophical concepts expressed in a domain thesaurus. The only requirement is that the annotations must be based on a "Annotation Domain Ontology" that is both loaded into the annotation tool and into Talia. More details about the Discovery projects are available in its website[10].

8. Item-centric vs relation-centric perspectives

In scholarly environments, where Talia is mostly expected to be used, the contex in which each individual object is placed is of extreme importance. It is often by exploring the context, that is the relations that each objects has with others, that new discoveries are made. As an example, consider the following figure.

The interface shown above is part of Hyper, the software Talia derives from. A similar interface is currently being re-implemented in Talia. It has been designed to visualize a particular type of relations that a set of manuscripts have among themselves. In particular, this interface allows the user to visualize a path of the genesis of a philosophical idea, from it's conception on the manuscript to its publication on a printed book, through its evolution and refinements in successive manuscripts and pre-printing copies. The following figure is an alternative visualization of one of the resources shown in the previous figure. This view, called

"rhizome view" shows all the "paths" that pass through a certain resource.

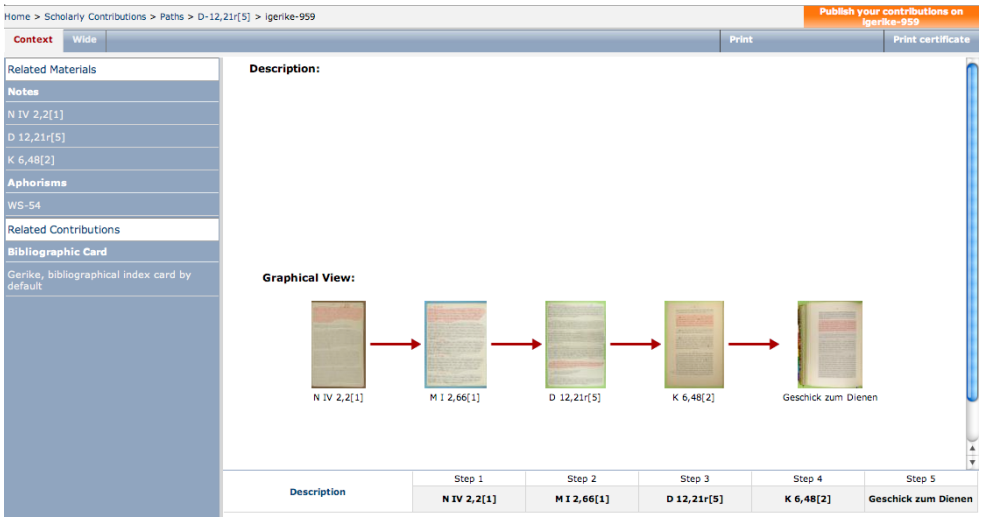


Figure 3: Path widget

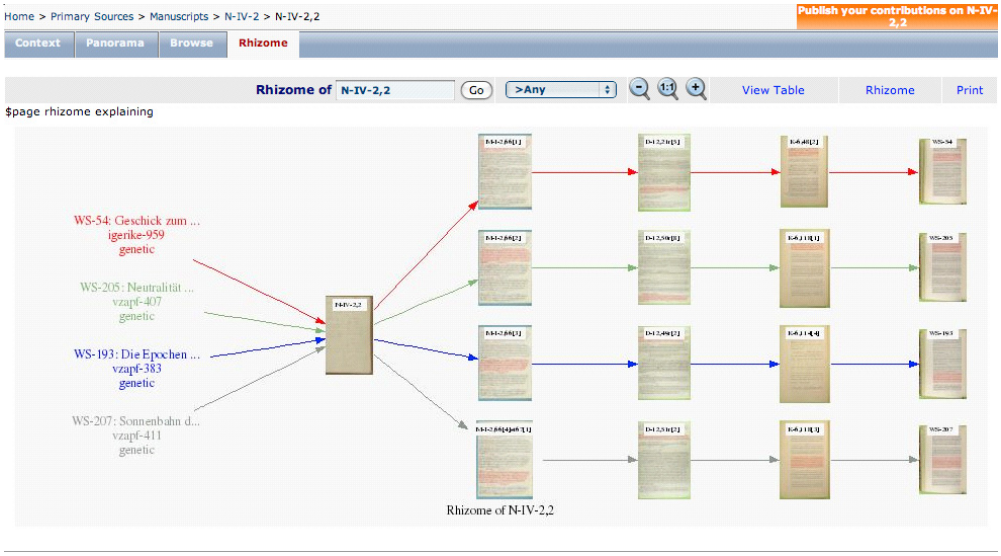


Figure 4: Rhizome widget

It is clear that the meaning of these two alternative visualizations is different from each other, but the interesting element is that in both these interfaces the focus is on relations among resources rather than on the resources themselves.

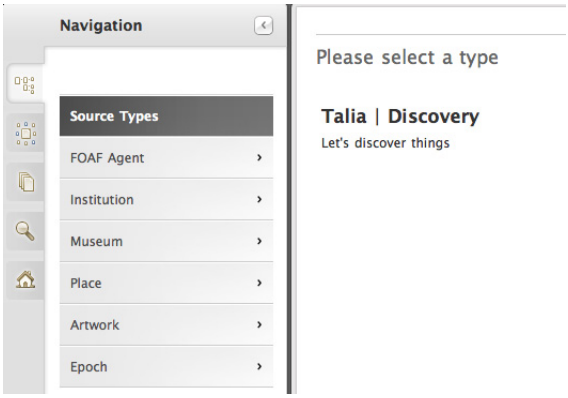
Having alternative interfaces that allow the user to focus on an individual resource as well as its relatioships with other resources is one of the charachteristics that makes Talia a scholarly tool rather than a simple digital object repository.

9. User Interface Widgets and Source Tool plugins

Talia is meant to be used to publish a vast amount of heterogeneous digital objects and resources. Scholarly communities are often interdisciplinary and their research output embraces more than one scientific domain. We believe, that general purpose user interfaces are unable to match the complexity of these contexts and

Figure 5: Default user interface. The Source is shown on the right hand side. The bar on the left lists semantic relations with other sources. Related sources are clickable.

fail to properly address the diverse needs of the users. Therefore, Talia provides a flexible and modular user interface framework based on widgets. Widgets are distributed independently and can be used as building blocks for customizing the application's user interface. Talia's Widgets engine offers an high level framework that can be used by application developers to build community specific user interfaces without the need of programming low level details. Widgets can easily be plugged into the default user interface.



The following figure shows an example of a Talia Semantic Navigation Interface, based on a widget that directly interacts with the Talia Knowledge Base, using metadata and ontologies to display information.

Figure 6: Semantic relations widget

We plan to host a library of Open Source widgets on Talia's website that developers can use to distribute their widgets.

In addition to the widgets component, Talia also includes another kind of plug-ins called Source Tools. These tools are behaviours that can be attached to a specific type of digital object (called "Sources" in Talia). For example a Source of type manuscript edition that includes a data object of type TEI-XML may have a Source tool that allows the user to perform some kind of linguistic analysis on the text. A source of type manuscript whose data objects are images representing the manuscript may have a Source Tool to OCR the the text from the image. The rizhome shown above is another example of Source Tool applied to a subset of Sources of type Manuscript.

10. Conclusions and Related works

Talia is an innovative distributed semantic digital library system, which aims at improving scholarly research in humanities by avoiding fragmentations of materials. Using standard Semantic Web technologies, ontologies to organize information and a completely customizable user interface framework, Talia represents a state-of-the-art research and publishing environment for humanities.

Talia is distributed with an Open Source license and it is very easy to install and configure so that it can be used to build single digital libraries and electronic publishing venues at a very low cost. Talia nodes, can then be joined together in a federation to create virtual collections that cross the borders of a single library or organization. At the same time, Talia's full compliance with Semantic Web standards ensures a deep interoperability with any other Semantic Web tools and applications.

Talia shares some properties with other semantic digital library systems like JeromeDL [11], BRICKS [12] and Fedora [13]. These projects are however mostly focused on the back-end technology and can hardly be deployed in low-tech environments such as small archives, libraries and museums. None of these tools offers tight coupling between the semantic knowledge base and the flexible user interface framework provided by Talia.

BRICKS is an architecture that is composed of a set of generic foundational components (called "core and basic services") plus a number of additional specialized services (called "Pillars") that can be invoked by applications as remote services. A BRICKS node (called Bnode) is an application that uses these services and interacts with other bNodes within a BRICKS network. BRICKS is therefore a huge infrastructures that requires a significant amount of central coordination to maintain the basic services. From the point of view of the individual institution that wants to join the network, BRICKS provides a set of very useful services on top of which each content provider should develop his own application and user interface.

We believe that within the Humanities and in general in the sector of cultural institutions, it is very uncommon that organizations have access to the know-how, budget and organizational capacity to deploy such a complex product. Moreover, even though the technology itself has a decentralized architecture, BRICKS relies on ad-hoc components ("core and basic services") that depend on the availability in the network of remote services. In this way the need of centralized coordination is shifted from the technological level to the organizational and managerial level.

We believe that the lack of organizational and managerial coordination is one of the weak spots of the Humanities scholarly community. Therefore, the approach of Talia is to minimize the efforts needed to set-up and deploy a Talia node. Additionally, a Talia federation does not rely on any legacy coordination and knowledge exchange protocol (such as the BRICKS P2P component). Talia is entirely built on top of very simple Semantic Web standards such as RDF, HTTP and URI's. In short, Talia is designed to run out-of-the-box in order to make it possible for smaller cultural institutions (whom may even not have or

have very small computing staff) to contribute to a Semantic Web of Culture.

The similarity between Talia and Fedora is that both allow to express relations between objects in RDF. However, "*...Fedora is a digital asset management (DAM) architecture, upon which many types of digital library, institutional repositories, digital archives, and digital libraries systems might be built. Fedora is the underlying architecture for a digital repository, and is not a complete management, indexing, discovery, and delivery application...*".

As with BRICKS, Fedora is suitable to develop and deploy very large digital library applications. Talia instead aims to be a complete out-of-the box application the comes with a pre-defined, generic and complete user interface. Talia also has a modular architecture that makes it easy to extend its features and customize its interface by developing plug-ins and UI widgets.

JeromeDL is the application most similar to Talia that currently exists. Like Talia, JeromeDL is fully based on simple Semantic Web Standards, works out-of-the-box and is extensible through plug-ins. Apart from the different language in which the two applications are written (JeromeDL is written in Java and Talia is written in Ruby) the main difference lies in their primary target user group. JeromeDL's primary target audience is the generic user of a digital library while Talia will also include default User Interfaces and tools that are targeted to Humanities Scholars.

At the time of writing, Talia is still in Alpha stage and a first stable public realease is planned for October 2008. The first release will include a set of visualization widgets specifically meant for handling Discovery content as well as a full-featured on-line peer review system. The first release will also include an adapter for Philospace, a semantic annotation tool based on the Dbin platform[14][15], briefly introduced in paragraph 7. In the meantime, Talia is being customized for other applications in the cultural heritage and digital library domains. In particular, additional research is being performed on the integration of Semantic Web based bibliometric tools. The focus of these research activity is to exploit the Semantic Web to explore bibliometric models and impact measures that can be used as alternative to traditional impact indicators such as the Impact Factor. Some of this models have been proposed in [16] and [17]. Other areas of future improvement include the development of an infrastructure for collaborative ontology editing and mapping as well as an ontology library for the Humanities and Cultural Heritage domains.

Finally, we are also studying the integration of Talia with archival cataloguing software and standards, such as EAD editors and archival data management systems, with the objective of making Talia a suitable product to interlink heterogeneous data and resources coming from the library, archival and museum domains to create digital collections of cultural resources.

11. Acknowledgements

This work has been supported by Discovery, an ECP 2005 CULT 038206 project under the EC eContentplus programme.

12. Notes

1 http://www.w3.org/TR/rdf-schema/

2 http://www.w3.org/TR/owl-features/

13. References

[1] The European Library Portal, [http://www.theeuropeanlibrary.org/portal/index.html]

[2] OAIster, [http://www.oaister.org/]

[3] W3C Semantic Web Activity, [http://www.w3.org/2001/sw/]

[4] RDF Primer, W3C Recommendation, [http://www.w3.org/TR/rdf-primer/]

[5] Nucci, M., David, S., Hahn, D., Barbera, M., *Talia: A Framework for Philosophy Scholars*, in proceedings of Semantic Web Applications and Perspective, Bari, Italy, 2007.

[6] Talia Wiki, [http://trac.talia.discovery-project.eu/]

[7] Fielding, R.T., *Achitectural Styles and the Design of Network-based Software Architectures*, PhD thesis, UC Irvine, 2000.

[8] OpenID Web Site, [http://openid.net/]

[9] Shibboleth Web Site, [http://shibboleth.internet2.edu/]

[10] Discovery Web Site, [http://www.discovery-project.eu/]

[11] Kruk, S., Woroniecki, T., Gzella, A., Dabrowski, M., McDaniel, B., *Anatomy of a social semantic library*, in: European Semantic Web Conference, Volume Semantic Digital Library Tutorial, 2007.

[12] Risse, T., Knezevic, P., Meghini, C., Hecht, R., Basile, F., *The bricks infrastructure - an overview*, in The International Conference EVA, Moscow, 2005.

[13] Fedora Development Team, *Fedora open source repository software: White paper*, white paper, Fedora Project, 2005.

[14] G. Tummarello, C. Morbidoni, M. Nucci, "Enabling Semantic Web communities with DBin: an overview", Proceedings of the Fifth International Semantic Web Conference ISWC 2006, Athens, GA, USA, 2006

[15] Dbin Web Site, [http://www.dbin.org/]

[16] Bollen, J., Van de Sompel, H., Smith, J., Luce, R., Towards alternative metrics of journal impact: a comparison of dowload and citation data. Information Processing & Management, Volume 41, Issue 6, pp. 1419- 1440, Dec. 2005

[17] Barbera, Michele and Di Donato, Francesca (2006) Weaving the Web of Science : HyperJournal and the impact of the Semantic Web on scientific publishing. In Martens, Bob and Dobrova, Milena, Eds. Proceedings ELPUB : International Conference on Electronic Publishing, pp. 341-348, Bansko Bulgaria, 2006.