Empowering Archives through Annotations

Nicola Ferro and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. The paper presents an integration and visualization service to enhance the use of annotations and to empower the role of the user and research community in the archival context. We show how this service allows us to address the interoperability between diversified digital archive and annotation systems. Furthermore, it propels the use of annotations to enhance the user experience and to exploit the archivists expertise both in the description and consultation phases.

1 Motivation

One of the main goal of the research on *Digital Library (DL)* is to supporting the creation of innovative applications and services to access, share and search our cultural heritage.

An important challenge in this field is to transform DL into a new type of information infrastructure that can be user-centered and able to support content management tasks together with tasks devoted to communication and cooperation [11]. DL can enable the intellectual production process and support user cooperation and exchange of ideas; in this way, DL not only foster access to knowledge, but they are also part of knowledge creation and evolution. The evolution and transmission of knowledge has always been an interactive process between scientists or field experts, and annotations have been one of the main tools for this kind of interaction. In the digital era, annotations are still a relevant means of intellectual collaboration and thus, one of the main collaboration tools exploited by DL [5].

The informative context enclosed by digital libraries is multifaceted and comprises many realities of interest such as libraries, archives and museums. In this paper we focus on the archives and archival metadata which are the basic means for accessing and consulting archival resources in a digital environment [18]. Annotations foster collaboration between archivists, researchers and general users by playing a central role both in the phase of *creation* and in the phase of *consultation* of archival metadata. Indeed, in the creation phase archivists have to select and describe the archival material and annotations allow them to explain and discuss their choices enabling users to properly access and consult the archival metadata. In the consultation phase, annotations are exploited to find out relationships between different parts of an archive or between different archives; for instance, users can exploit annotations to move from one archive to another guided by the expertise of the archivists that annotated them. In order to properly exploit annotations in the archival context we have to take

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 57–68, 2013.

into account the heterogeneous environment composed of digital archive systems and annotation systems which are often grounded on different methodological and technological approaches. The archival community has developed "content and data structure standards" [15] to facilitate the description, management and access to the archival resources; however, these standards can be difficult for archivists to use [19] and are often implemented in ways that can negatively affect their description activity [20]. Thus, there has been a proliferation of digital archival systems based on diversified descriptive methodologies and metadata; also from the annotation point-of-view a lot of research has been done that has led to the design and development of variegated annotation systems [4].

This heterogeneity turns into an interoperability problem when we need to access and consult archival metadata managed by different digital archive systems and annotations created and handled by different systems. On the other hand, every digital archive system has to respect some fundamental archival principles – i.e. the hierarchical organization of the documents and their descriptions [6]; moreover, also annotations under certain conditions can be opportunely organized in a hierarchical way [4]. We exploit these facts to define a common basis for addressing interoperability issues and for designing an integration and visualization service for annotated archives. To this end we use the *NEsted SeTs for Object hieRarchies (NESTOR)* Model [8] and the *Flexible Annotation Service Tool (FAST)* annotation model [4] to:

- propose a methodology which provides us with a unified, coherent, and concise view of heterogeneous archival metadata and annotations;
- design a service allowing users to consult different archives within the relative annotations and find out the relations between different archives connected by annotations;
- develop a Web-based visualization tool based on this service which helps users to access and consult archival metadata and annotations.

The paper is organized as follows: Section 2 gives a brief background about archives, archival metadata, and annotation services highlighting the concepts we exploit in the rest of the work. Section 3 reports on the heterogeneity of archival metadata. Section 4 presents the methodology which by using the NESTOR Model and the FAST annotation model allows us to represent archives and annotations in an integrated and coherent way. Section 5 describes the proposed architecture of the integration and visualization service. In Section 6 we present the functioning of the Web-based visualization tool prototype. Finally, in Section 7 we conclude and present some future works.

2 Background

Archives. An archive is the trace of the activities of a physical or juridical person in the course of their business which is preserved because of their continued value. Archives have to keep the context in which their records have

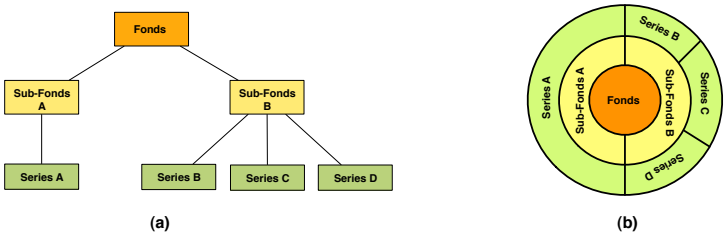

Fig. 1. The structure of a sample archive represented by: (a) a tree; (b) a Doc-Ball

been created and the network of relationships between them in order to preserve their informative content and provide understandable and useful information over time [10]. The context and the relationships between the documents are preserved thanks to the hierarchical organization of the documents inside the archive. Indeed, an archive is divided by fonds and then by sub-fonds and then by series and then by sub-series and so on – see Figure 1a for an example; at every level we can find documents belonging to a particular division of the archive or documents describing the nature of the considered level of the archive (e.g. a fonds, a sub-fonds, etc.). The union of all these documents, the relationships and the context information permits the full informational power of the archival documents to be maintained. The archival documents are analyzed, organized, and recorded by means of the *archival descriptions* [12] that have to reflect the peculiarities of the archive [6].

Digital Archives and the NESTOR Model. In the digital environment archival descriptions are encoded by the use of metadata; these need to be able to express and maintain the structure of the descriptions and their relationships [10]. Archives can be modeled by means of the NESTOR Model which relies on two set data models called *Nested Set Model* (NS-M) and *Inverse Nested Set Model* (INS-M) [3]. Both these set data models, formally defined in the context of set theory, can be used to model an archive by means of nested sets [8]. An extensive analysis of the NESTOR Model and its applications in the context of DL and archives can be found in [3]; in this paper we exploit the functionalities of the INS-M and thus we focus our presentation on this model.

The most intuitive way of understanding how the INS-M works is to see how a sample tree is mapped into an organization of nested sets based on the INS-M. We can say that a tree is mapped into the INS-M by transforming each node into a set, where each parent node becomes a subset of the sets created from its children. The set created from the tree's root is the only set with no subsets and the root set is a proper subset of all the sets in the hierarchy. The leaves are the sets with no supersets and they are sets containing all the sets created from the nodes composing the tree path from a leaf to the root. We can represent in a straightforward way the INS-M by means of the "*DocBall representation*" [17] – see Figure 1b. It is worthwhile to understand how the DocBall is used because the graphical tool we are going to present is based on this idea. The DocBall

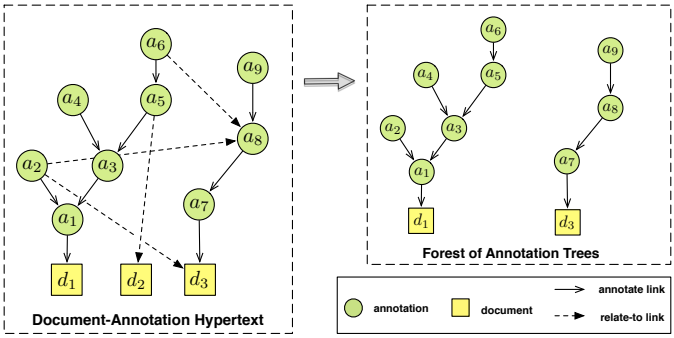

Fig. 2. A document-annotation hypertext and its subgraph composed of the "annotate" edges which is a forest composed of two trees

is composed of a set of circular sectors arranged in concentric rings; each ring represents a level of the hierarchy with the center representing the root. In a ring, the circular sectors represent the nodes in the corresponding level. We use the DocBall to represent the INS-M, thus for us each circular sector corresponds to a set; for instance, referring to Figure 1b, it is possible to say that section "Series C" is a direct superset of section "Sub-Fonds B".

Annotations and the FAST Annotation Model. Research on annotations has given rise to different data models, systems and services. An example is the MPEG-7[1] which is a standard for annotating and describing multimedia content data; the Semantic Web is another example of where annotations are exploited, in particular in the context of the Annotea project developed by the W3C[2]. In the context of DL an example is *Collaboratory for Annotation Indexing and Retrieval of Digitized Historical Archive Material* (COLLATE) [16], which supports the collaboration among film scientists and archivists.

Another relevant example is FAST which adopts and implements the formal model for annotations proposed in [4]. FAST distinguishes between *documents* – which are generic digital objects managed by a DL – and *annotations*. Annotations can be associated with a digital object by two types of link: **annotate link** and **relate-to link**. An annotate link allows an annotation to be linked to a part of a digital object; through this link it is possible to express *intra-digital object relationships* between different parts of an object. A relate-to link is intended to allow an annotation only to relate to one or more parts of other digital objects, but not the annotated one; therefore it expresses *inter-digital object relationships*. From these definitions annotations can be seen as a means of linking digital objects. Annotations permit us to create new relationships between the components of a digital object, between different digital objects of the same DL or between digital objects belonging to different DL. As shown in [4] the set

[1] Please refer to *ISO/IEC 15938-1:2002*.
[2] `http://www.w3.org/2001/Annotea/`

of digital objects and annotations form a labeled directed acyclic graph called document-annotation hypertext. Furthermore, each annotation must annotate only one digital object, and it has been shown [4] that for each document there is a **unique tree of annotations** constituted by "annotate" edges that can be rooted in the document. In Figure 2 we can see an example of document-annotation hypertext and the trees formed by the "annotate" links.

3 Heterogeneity of Archival Metadata

The standard format of metadata for representing the hierarchical structure of the archive is the *Encoded Archival Description (EAD)* [13], which reflects the archival structure and holds relations between entities in an archive. In addition, EAD has a flexible structure, encourages archivists to use collective and multilevel description, and has a broad applicability. On the other hand, the EAD permissive data model may undermine the very interoperability it is intended to foster and it must meet stringent best practice guidelines to be shareable and searchable [15]. Furthermore, an archive is described by means of a unique EAD file and this may be problematic when we need to access and exchange archival metadata with a variable granularity [7] by means of DL standard technologies like the *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*[3].

Although EAD is the archival description standard, several other modeling methodologies and metadata formats have been developed. Indeed, we may consider the "Tree-based Metadata" approach in which archives are described by a collection of lightweight metadata – e.g. Dublin Core Application Profiles[4] – one for each archival resource, connected to each other by means of links to a third-party file – e.g. an external XML file – which maintains the archival structure [14]; alternative instantiations of this approach maintain the archival structure by means of an opportunely designed relational database [15]. These approaches differ from EAD both in the way in which they express the structure and the content of the archive. Furthermore, outside EAD boundaries, there is no common agreement on which metadata fields should be used to describe archival resources.

There is also the possibility of representing the archival structure by means of the INS-M [7]. It has been shown [8] that an archive can be modeled by means of the INS-M and then instantiated in such a way that allows the use of the OAI-PMH architecture to enable a variable granularity access and exchange of the archival metadata. Furthermore, [7] describes a methodology to map an EAD file into the NESTOR Model and preserve the full informative power of the metadata. Mapping an EAD file into the NESTOR Model means that we make use of a methodology that maps the EAD structure into the INS-M and a collection of lightweight metadata containing the content information retained by EAD. In this way the INS-M preserves the archival structure while the metadata

[3] `http://www.openarchives.org/`
[4] `http://www.dublincore.org/`

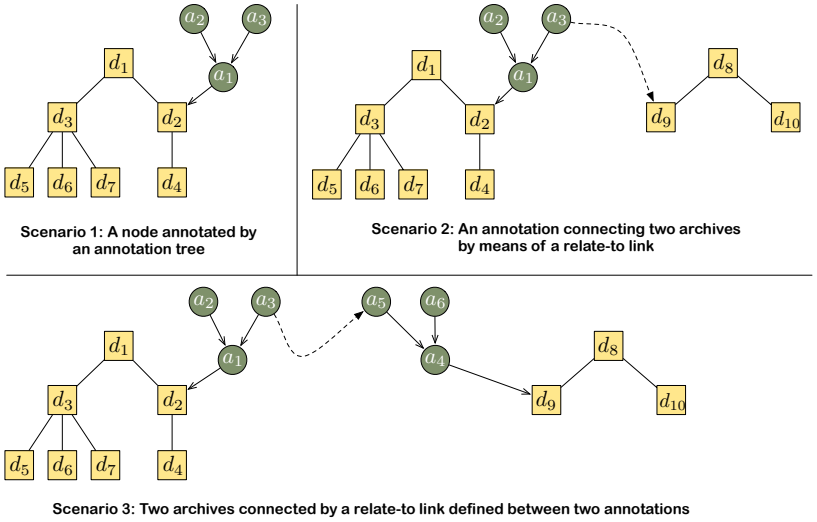

Fig. 3. Annotations: Three possible scenarios in the archival context

belonging to its sets preserve the content of archival descriptions [7]. In the same way, this methodology is adopted with the "Tree-based metadata" approach, where the structure retained by an external XML file or by a relational database is mapped into the INS-M [3].

4 An Integrated View of Archives and Annotations

Our goal is to make available a uniform and integrated view of archives described and managed by means of heterogeneous digital archive systems together with their annotations which in turn can be handled by different annotation systems. To this purpose we rely on the NESTOR Model and on the FAST annotation model to address interoperability at the archival level and to show how annotations can be enclosed in the "NESTOR view" of the archives.

We present three possible scenarios showing how annotation trees can be attached to an archive and then we show how they can be modeled through the INS-M and represented by means of the DocBall. Figure 3 presents the scenarios; in this figure an archive is represented as a document tree where the nodes are named as "$d_1, d_2, \ldots$" for convenience; for the same reasons annotations are indicated as "$a_1, a_2, \ldots$". In the first scenario we consider an archival tree where the node d_2, annotated by a_1, is the root of an annotation tree composed of three annotations. The second scenario shows that a_3 which is part of an annotation tree annotating d_2 is connected to a second archive by means of a "relate-to" link. In the third scenario, we can see two archives connected by a relate-to link defined between two annotations – i.e. a relate-to link between a_3 and a_5. Figure 4 shows by means of the DocBall representation how these scenarios are handled by the INS-M; we adopt the DocBall as a graphical means to describe

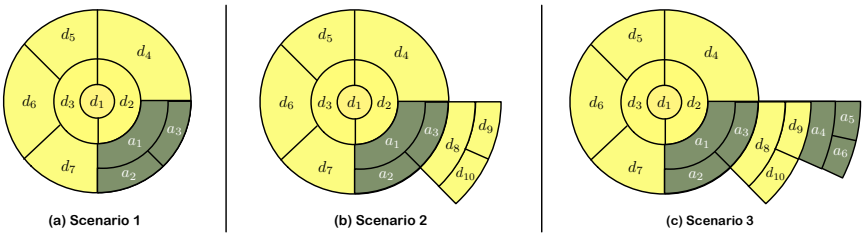

Fig. 4. DocBall representations of archive annotation scenarios

and explain how archives and annotations are joined together by means of the INS-M.

In the first scenario we need to join an "archival DocBall" representing the archive and an "annotation DocBall" representing the annotation tree originally attached to node d_2 of the archive – see Figure 3a. The resulting DocBall is shown in Figure 4a, where a_1 is a superset of d_2. The second scenario presents the same annotated archive we have seen in the first scenario enriched by the relationship of annotation a_3 with the node d_9 of a second archive. In this case, we use a DocBall representing the first archive within its annotations – call it "DocBall A" (see Figure 4a) – and a DocBall representing the second archive – call it "DocBall B". In order to join these two DocBalls connected by annotation a_3, we add the inner sector of DocBall B – i.e. d_8 – to DocBall A as a superset of a_3. The resulting DocBall (see Figure 4b) provides us with of an integrated view of the two archives connected by the annotation tree rooted in a_1. The third scenario enhances this idea; indeed, in this case both "DocBall A" and "DocBall B" represent annotated archives that have to be joined together. So, we follow the methodology presented for scenario 2 by taking the inner sector of DocBall B – i.e. d_8 which represents the root of the second archive – and adding it to DocBall A as a superset of the annotation – in this case a_3 – which relates the two archives together. The general methodology of joining two DocBall together can be summarized as follows; let D_A and D_B be two DocBall, where section s_A of D_A is *related to* section s_B of D_B. To join D_A with D_B, the inner section of D_B must be added to D_A as a superset of s_A.

5 Architecture and Functionalities of the Integration and Visualization Service

In order to accomplish the purposes of this work, the integration and visualization service must be non-intrusive, scalable and flexible. Indeed, it has to be *non-intrusive* to model the archives and annotations by means of the INS-M without interfering with the organization and the functioning of the local systems. It has to be *scalable* to collect resources in a distributed environment, and it has to be *flexible* to integrate archives and annotations together satisfying user needs.

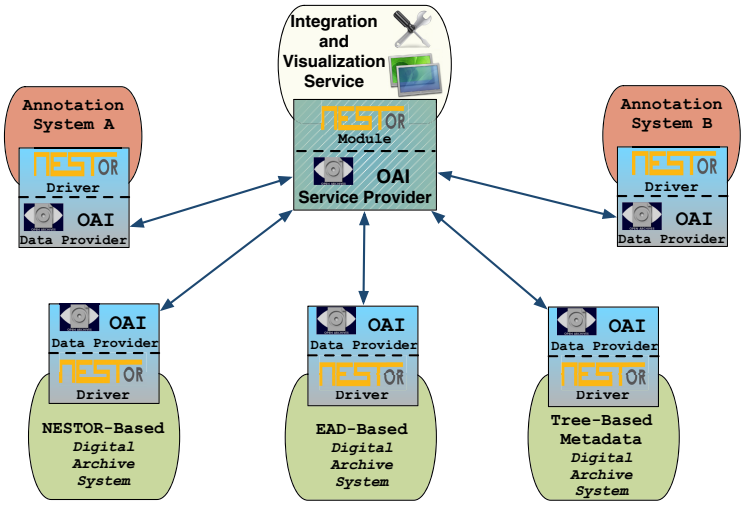

Fig. 5. Proposed architecture of the integration and visualization service

In Figure 5 we can see the proposed architecture of the integration and visualization service. We consider three different digital archive systems: the first is based on the NESTOR model, the second on EAD and the third on the "Tree-based metadata" approach. Furthermore, we consider two generic annotation systems: "Annotation system A and B". Each digital archive and annotation system should be equipped with a software module divided into two main components. The first component is called "NESTOR driver" and the second is an OAI Data Provider. The NESTOR driver is a lightweight component that has to map the archival metadata into the INS-M and prepares them to be exchanged by means of OAI-PMH. If we consider the NESTOR-based system in Figure 5, the NESTOR driver has to check if the archival metadata are modeled by means of NS-M or INS-M and in the first case it has to map the archive from the NS-M into the INS-M [8]. For the EAD-based archive system, the NESTOR driver has to map the EAD files into the INS-M [7] and in the "Tree-based metadata" system it has to map the XML file or the relational schema preserving the archive structure into the INS-M [3]. The NESTOR driver does the same operations with the annotation trees by mapping them into the INS-M [9].

In this way the NESTOR driver addresses the heterogeneity between different digital archive and annotation systems in a non-invasive and transparent way: the local systems handle archives and annotations within their own policies and expose them coherently with the INS-M. Furthermore, we know that the sets and the metadata defined by the INS-M can be straightforwardly exchanged by means of OAI-PMH [8]. Thanks to this feature we can exploit the OAI-PMH architecture to exchange the archival metadata and the annotations between the local systems and the centralized integration and visualization service. As we can see in Figure 5, the NESTOR driver is configured as a component of an

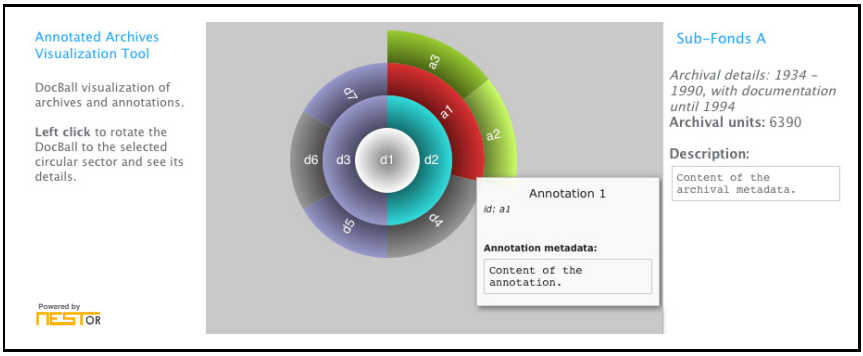

Fig. 6. Prototype of the visualization Tool: DocBall representation of Scenario 1

OAI Data Provider; in this way the presented architecture draws on OAI-PMH scalability and flexibility and the NESTOR driver can be configured as a plug-in of already existing and widely-diffused software modules.

The integration and visualization service can be developed over an OAI Service Provider which harvests the archival metadata and the annotations. The service utilizes a "NESTOR module" that acts as a mediator between the requests of the service and the harvested metadata and annotations. According to the three scenarios presented in the previous section, if the service requires just the archival metadata together with their annotations – i.e. scenario 1 – the NESTOR module embeds the archive with its annotations and returns the INS-M represented by the DocBall in Figure 4a. The NESTOR module returns a DocBall like the ones in Figure 4b and 4c when the service needs to exploit the relationships established by the annotations between different archives. The role of the visualization tool is to enhance the relationships between an archive and its annotations and between different archives connected by annotations. Especially in the second and third scenarios, the visualization tool needs to have an effective interface to help the users to infer and exploit the relationships between the resources.

6 Web-Based Visualization Tool

The visualization tool is the front-end component of the integration and visualization service; it relies on the archives and annotations modeled by means of the INS-M. We show and discuss several screenshots of the initial prototype of the visualization tool based on test data; Figure 6 shows how the service addresses the first scenario. We can see that the DocBall is similar to the one in Figure 4a and it shows an archive where section d_2 is annotated by an annotation tree composed of three annotations. In the left column we have general information about the service. The DocBall is in the center of the canvas and when we move the pointer over a circular section a tooltip appears showing the content of this section; if we click on a section, the DocBall rotates and the selected section is

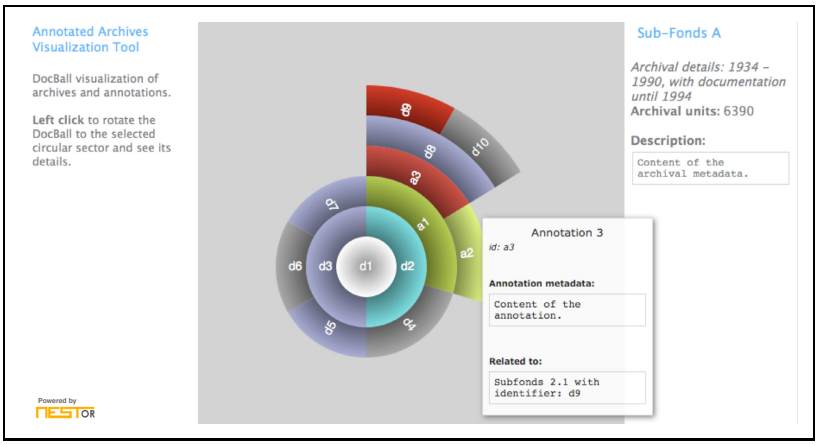

Fig. 7. Prototype of the visualization Tool: DocBall representation of Scenario 2

highlighted. In this figure we selected section d_2 the content of which is shown in the right column and the tooltip shows the content of a_1. In this way the user can select an archival section, see its content in the right column and view the content of annotations or other archival divisions by means of the tooltip.

Figure 7 shows a screenshot where the visualization tool addressed the second scenario; the annotation (a_3) which annotates an archival section (d_2) is related to the archival section (d_9) of a second archive. The tool highlights the related sections; indeed, when d_2 is selected, the DocBall rotates in such a way that its annotation tree moves on the top of the DocBall, and annotation a_3 together with section d_9 are colored in red reveling the connection between them. The user can explore the content of these sections by means of the tooltip while visualizing the content of d_2 in the right column. In Figure 7 we captured the tooltip related to a_3; we can see that it reports the content of the annotation and the information about its relationship with section d_9.

Figure 8 shows the last scenario where we exploit the relationship between two annotations – i.e. a_3 and a_5 – to relate two different archives. The service works as in the second scenario but in this case it highlights the two annotations; the user can visualize the content of the annotations of the first and second archive as well as the content of the second archive contextually with the content of the selected archival section.

We can see that archival documents and annotations are represented as circular sectors with different colors in the DocBall. The use of colors may be an effective way to distinguish between the sectors which are documents and those which are annotations. Furthermore, the DocBall could become ineffective if there are many sectors that have to be represented. In this case an expand/compress strategy could be adopted as well as it is used to shows the branches of very large trees.

Fig. 8. Prototype of the visualization Tool: DocBall representation of Scenario 3

7 Conclusion

In this paper we propose the architecture of an integration and visualization service that exploits the NESTOR Model and the FAST annotation model to provide us with a unified view of archives and annotations that can come from diversified systems. This service can address interoperability issues between different digital archive systems and annotation systems in a flexible and scalable way by exploiting existing and widely-diffused software modules – i.e. OAI Data and Service Provider – and extending them by means of lightweight software modules – i.e. the NESTOR driver. The presented prototype of the service enables a comprehensive view of archival structure and content together with its annotations; furthermore, it highlights the relationships between different archives. This service can enhance the role of annotations in the archival context and the expertise of archivists in the description as well as in the consultation phase of the archives.

Future work foresees the adoption of this service in the context of a project of the Italian Veneto Region[5]. The main aim of the project is to make available a regional archival information system which allows the management of the resources of archives present in the Region.

Acknowledgments. The work reported has been carried out in the context of an agreement between the Italian Veneto Region and the University of Padua. CULTURA[6] (Grant agreement no. 269973) and the PROMISE network of excellence[7] (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

[5] http://www.regione.veneto.it/
[6] http://www.cultura-strep.eu/
[7] http://www.promise-noe.eu/

References

1. Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.): ECDL 2008. LNCS, vol. 5173. Springer, Heidelberg (2008)
2. Agosti, M., Esposito, F., Thanos, C. (eds.): IRCDL 2010. CCIS, vol. 91. Springer, Heidelberg (2010)
3. Agosti, A., Ferro, N., Silvello, G.: The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures. In: Proceedings of the 18th Italian Symposium on Advanced Database Systems, pp. 242–253. Società Editrice Esculapio, Bologna (2010)
4. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Trans. Inf. Syst. 26(1) (2007)
5. Agosti, M., Ferro, N.: Annotations: A Way to Interoperability in DL. In: [1], pp. 291–295
6. Duranti, L.: Diplomatics: New Uses for an Old Science. Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, Lanham, Maryland, USA (1998)
7. Ferro, N., Silvello, G.: A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In: [1], pp. 268–279
8. Ferro, N., Silvello, G.: The NESTOR Framework: How to Handle Hierarchical Data Structures. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 215–226. Springer, Heidelberg (2009)
9. Ferro, N., Silvello, G.: FAST and NESTOR: How to Exploit Annotation Hierarchies. In: [2], pp. 55–66
10. Gilliland-Swetland, A.J.: Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment. Council on Library and Information Resources, Washington, DC (2000)
11. Kani-Zabihi, E., Ghinea, G., Chen, S.Y.: Experiences with Developing a User-Centered Digital Library. IJDLS 1(1), 1–23 (2010)
12. Pearce-Moses, R.: Glossary of Archival And Records Terminology. Society of American Archivists (2005)
13. Pitti, D.V.: Encoded Archival Description. An Introduction and Overview. D-Lib Magazine 5(11) (1999)
14. Prom, C.J., Habing, T.G.: Using the Open Archives Initiative Protocols with EAD. In: Proc. 2nd ACM/IEEE Joint Conf. on Digital Libraries, pp. 171–180. ACM Press, USA (2002)
15. Prom, C.J., Rishel, C.A., Schwartz, S.W., Fox, K.J.: A Unified Platform for Archival Description and Access. In: Proc. 7th ACM/IEEE Joint Conf. on Digital Libraries, pp. 157–166. ACM Press, USA (2007)
16. Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A., Neuhold, E.J.: COLLATE - A Collaboratory Supporting Research on Historic European Films. Int. J. on Digital Libraries 4(1), 8–12 (2004)
17. Vegas, J., Crestani, F., de la Fuente, P.: Context Representation for Web Search Results. Journal of Information Science 33(1), 77–94 (2007)
18. Vitali, S.: Archival Information Systems in Italy and the National Archival Portal. In: [2], pp. 5–11
19. Yakel, E., Shaw, S., Reynolds, P.: Creating the Next Generation of Archival Finding Aids. D-Lib Magazine 13(5/6) (May/June 2007)
20. Yako, S.: It's Complicated: Barriers to EAD Implementation. American Archivist 71(2) (Fall/Winter 2008)