Enhancing Digital Cultural Heritage Collections with Social Network Capabilities

Maristella Agosti
Dept. of Information Engineering
University of Padua
Via Gradenigo, 6/a
Padua, Italy
agosti@dei.unipd.it

Nicola Ferro
Dept. of Information Engineering
University of Padua
Via Gradenigo, 6/a
Padua, Italy
ferro@dei.unipd.it

Sara Porat
Dept. of Information & Social Analytics
IBM Haifa Research Lab,
Haifa University,
Mount Carmel, Haifa 31905, Israel
porat@il.ibm.com

Ella Rabinovich
Information Retrieval Group
IBM Haifa Research Lab,
Haifa University,
Mount Carmel, Haifa 31905, Israel
ellak@il.ibm.com

ABSTRACT

This paper discusses how annotations can be exploited to increase the engagement of researchers with digital cultural heritage collections by inferring a social network among researchers from their annotations and discovering implicit relationships among them.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *collection, dissemination, systems issues, user issues.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services - *data sharing, Web-based services.*

General Terms

Algorithms, Management, Design, Experimentation, Human Factors.

Keywords

Cultural heritage collections, digital cultural heritage collections, digital libraries and archives, data curation, annotation, digital humanities, personalization, recommendation, entity extraction.

1. INTRODUCTION

Almost everybody is familiar with annotations and has his own intuitive idea about what they are, drawn from personal experience and the habit of dealing with some kind of annotation in everyday life, which ranges from jottings for the shopping to taking notes during a lecture or even adding a commentary to a text [2,3]. This intuitiveness makes annotations especially appealing for both researchers and final users: the former propose annotations as an easy understandable way of performing user tasks, while the latter feel annotations to be a familiar tool for carrying out their own tasks. Therefore, annotations have been adopted in a variety of different contexts, such as content enrichment, data curation, collaborative and learning applications, and social networks, as well as in various information management systems, such as the Web (semantic and not), digital libraries, and databases.

The CULTURA environment [9,10] is supporting different use cases, one of which are the 1641 Depositions[1], a collection of noisy text documents, mainly of a legal nature, dating from the 17th Century. They primarily contain witness testimonies from Protestants, but also some Catholics, from all social backgrounds. The collection, which has been digitized and transcribed, contains over 8,500 depositions of 20,000 pages, in which men and women of all classes and from all over Ireland told of their experiences following the outbreak of rebellion by the Catholic Irish in October 1641. This body of material provides a unique source of information for the causes and events surrounding the 1641 rebellion and for the social, economic, cultural, religious, and political history of seventeenth-century Ireland, England and Scotland. This is typical of the category of digital resource which will benefit most from CULTURA as it is inconsistent in spelling, punctuation, nomenclature and word forms, and reflects a cultural outlook quite different to the modern one.

The CULTURA environment allows researchers and historians to cooperate together and to add annotations[2]. In particular, in this paper, we are interested in exploring how annotations can be exploited as a vehicle for fostering a social network of researcher, how they can put researchers in contact and how analysis and mining on their structure and content can help in discovering and inferring relationships among them.

The paper is organized as follows: Section 2 provides a brief use case to exemplify the exploitation of annotations we are aiming at. Section 3 describes how we plan to extend the current entity model of the CULTURA project to support social services among researchers. Finally, Section 4 draws some conclusions and provides an outlook for future work.

[1] http://1641.tcd.ie/index.php

[2] http://cultura-project.eu/

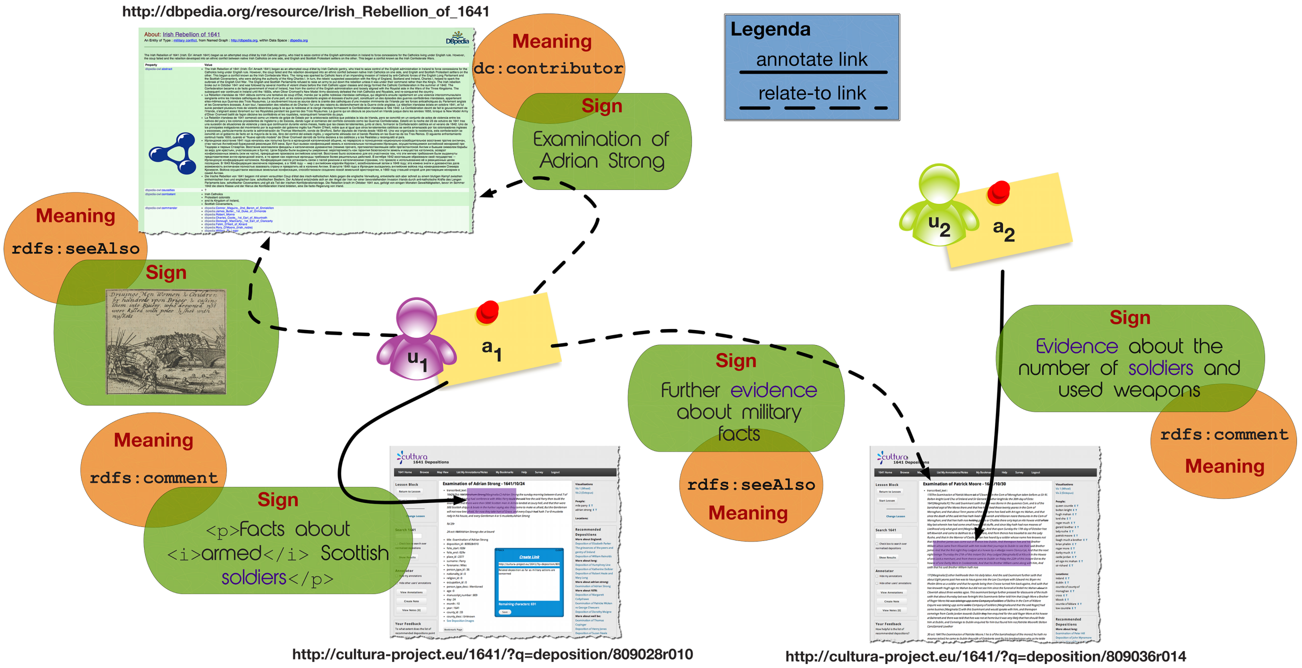
http://dbpedia.org/resource/Irish_Rebellion_of_1641
Meaning
dc:contributor
Sign
Examination of Adrian Strong
Legenda
annotate link
relate-to link
Meaning
rdfs:seeAlso
Sign
u_2
a_2
u_1
a_1
Sign
Evidence about the number of soldiers and used weapons
Sign
Further evidence about military facts
rdfs:seeAlso
Meaning
rdfs:comment
Meaning
Meaning
rdfs:comment
Sign
<p>Facts about <i>armed</i> Scottish soldiers</p>
http://cultura-project.eu/1641/?q=deposition/809028r010
http://cultura-project.eu/1641/?q=deposition/809036r014

Figure 1. Example of annotation.

2. USE CASE

The Flexible Annotation Semantic Tool (FAST) service adopts and implements the formal model for annotations proposed by [4] which has been also embedded in the reference model for digital libraries developed by DELOS, the European network of excellence on digital libraries [6]. It provides the back-end for the annotation functions offered by the CULTURA environment [1].

According to this model, an annotation is a compound multimedia object that is constituted by different signs of annotation. Each sign materializes part of the annotation itself; for example, we can have textual signs, which contain the textual content of the annotation, image signs, if the annotation is made up of images, and so on. In turn, each sign is characterized by one or more meanings of annotation, which specify the semantics of the sign; for example, we can have a sign whose meaning corresponds to the title field in the Dublin Core (DC) metadata schema[3], in the case of a metadata annotation, or we can have a sign carrying a question of the author's about a document whose meaning may be "question" or similar.

An annotation has a scope which defines its visibility (public, shared, or private), and can be shared with different groups of users. Public annotations can be read by everyone and modified only by their owner; shared annotations can be modified by their owner and accessed by the specified list of groups with the given access permissions, e.g. read only or read/write; private annotations can be read and modified only by their owner.

The flexibility inherent in the annotation model allows us to create a connective structure, which is superimposed to the underlying documents managed by digital libraries. This can span and cross the boundaries of different digital libraries and the Web, allowing the users to create new paths and connections among resources at a global scale.

Figure 1 shows an example of annotation which summarizes the discussion so far.

The annotation, with identifier a_1, is authored by the researcher u_1. It annotates the deposition by Adrian Strong, whose identifier is http://cultura-project.eu/1641/?q=deposition/809028r010. The annotation relates to another deposition by Patrick Moore, whose identifier is http://cultura-project.eu/1641/?q=deposition/809036r014; in addition, it relates also to the DBpedia page of Irish Rebellion of 1641, http://dbpedia.org/page/Irish_Rebellion_of_1641, where Adrian Strong was one of the witnesses.

In particular, a_1 annotates a region of the deposition by using a textual sign whose content is "Facts about *armed* Scottish soldiers" and whose meaning is to be a comment in the RDFS namespace, i.e. a comment according to the RDF Schema W3C recommendation [5]. Note how the content of the sign is HTML to allow for richer formatting. In general, the content of a sign is specified by its MIME media type and this allows for great flexibility and for embedding different formats, such as XML, RDF, and so on.

The annotation with identifier a_1 annotates the deposition by Adrian Strong to one by Patrick Moore, with a textual sign whose content is "Further evidence about military facts" and whose meaning is to seeAlso in RDFS. This annotation thus represents the outcomes of the actual work of historian, who conducted his/her own research on these two depositions, to determine that they provide joint evidence about military facts.

[3] http://dublincore.org/

Moreover, a1 relates the Adrian Strong deposition to the DPpedia page of Irish Rebellion of 1641, which is the context surrounding the deposition, with two signs: a textual sign whose content is "Deposition of Adrian Strong" and whose meaning is contributor in the Dublin Core metadata schema; and, an image sign with a picture taken from James Cranford, Teares of Ireland (London, 1642)[4], whose meaning is "see also" in the RDFS namespace.

The annotation with identifier a_2 is authored by the researcher u_2. It annotates the deposition by Patrick Moore, with a textual sign whose content is "Evidence about the number of soldiers and used weapons" and whose meaning is to be a comment in the RDFS namespace.

Annotations a_1 and a_2 can be exploited in several ways to create a social network between the researchers u_1 and u_2 and to infer relationships among them. The plain fact that they are annotating and referring to the same deposition can be used as an evidence of related interested between u_1 and u_2. The fact that this link is typed by the meaning of the signs of its annotations can be exploited to further refine the analysis and discovering of relationships among researchers. Finally, we can even take into consideration the actual content of the annotations, where even a simple similarity search may reveal common terms or topics, as highlighted in violet in Figure 1.

The next section described how the CULTURA entity model and search capabilities can be extended to support this vision and scenario. With respect to searching annotation, FAST offers a search mechanisms [11,12,13] focused on the annotations and annotated resources, which can be exploited as a facilitator for analyzing annotation in the extended CULTURA entity model.

3. EXTENDING CULTURA ENTITY MODEL TO SUPPORT SOCIAL SERVICES

Carmel et. al. [7] describe an entity oriented search and exploration system that was developed for the EU CULTURA project. The system uses an extended faceted search solution for indexing and searching unstructured and semi-structured data sets, through an innovative approach where all entities and their relationships are searchable and retrievable. The solution, as described in Yogev et al. [14], utilizes an extension of the classical Entity-Relationship conceptual model to model data discovery requirements, and a logical document model for representing and indexing entity-relationship data.

The CULTURA Environment leverages a NLP module for extracting the key individuals, locations and events within cultural textual resources, and generating Entity-Relationship data set according to the schema depicted in Figure 2. Entity-oriented Search (EoS) is then used to search, explore and navigate over the depositions, the extracted entities and the relationships between them.

In this paper we suggest to go beyond entities that are mentioned in the depositions, and augment the Entity-Relationship schema with entities that represent a group of researchers that interact with the collection via FAST. Figure 3 shows an augmented schema that captures the group of researchers, and represents an annotation through a relationship between the researcher and the annotated deposition.

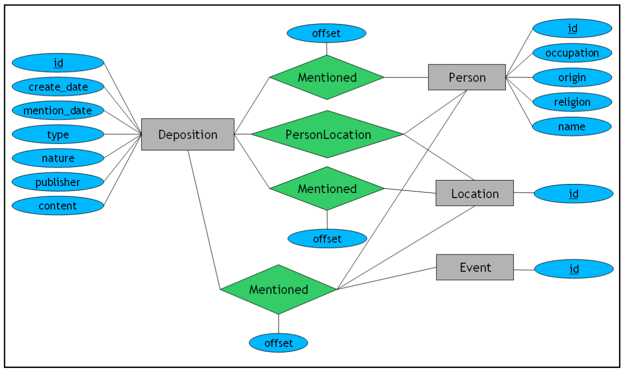

Figure 2. ER model for 1641 corpus.

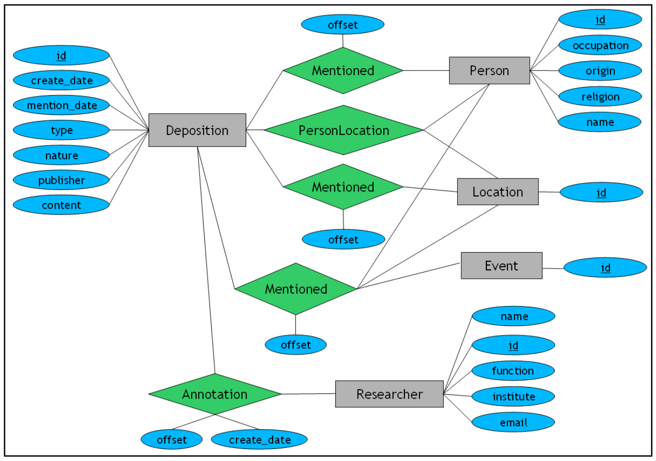

Figure 3. ER model for 1641 corpus including researchers and annotations.

We offer a way to construct a social network of researchers, based on *inferred* relationships. An inferred relationship between researchers involves co-annotation of a certain deposition, as depicted in Figure 4. The strength of the relationship takes into account the two attributes of the relationships involved. For instance, if the two different researchers work on different parts of the same deposition, the inferred relationship strength would be weaker than that of researchers working on the same parts of the deposition. It is possible to tell these two cases apart by comparing the offset attribute of both relationships – if the offsets are similar the relationship between the researchers would receive higher score.

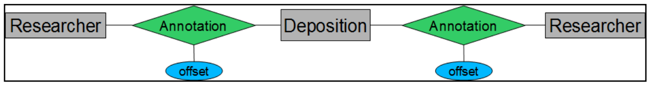

Figure 4: Relationship between two researchers co-annotating a certain deposition.

The social network between researchers can serve as the base for social services, such as personalized content recommendation and adaptive personalized search. As an example, each researcher can be assigned with a "personal profile" by locating a weighted list of the researchers close to her in the social network. Then, each time the researcher logs into the system, or periodically, the system will issue a query which identifies annotation relationship

[4] Image produced by the courtesy of the Board of Trinity College, Dublin and taken from: http://1641.tcd.ie/about.php

in which the researchers from the personal profile participate, and return these annotations or the relevant depositions. The relationships can be weighted according to the weight of researchers in the personal profile, and they can be further limited by creation date to ensure only new annotations are taken into account.

More recent research focused on locating topic experts in the organization [8]. In contrast to the previous service, in this case the focus is not on the social network of researchers, but rather on the relationships between researchers and content. The Cultura scenario where such a service would be needed is the case where a researcher first encounters a topic, either a term or an entity (person, location). The researcher may be able to utilize existing knowledge by finding colleagues who already worked on this topic. This requires identifying cases where the topic of interest is included in a text covered by an annotation. Given that all these cases are identified, the researchers who created the annotations may have knowledge on the subject, and they should be ranked according to the number of times each researcher created an annotation which covers the topic.

4. CONCLUSIONS

The paper presented several innovative ideas on how to exploit annotations to foster collaboration and relationships among researchers in digital humanities. Future work will concern the actual implementation of the proposed ideas into a system prototype and its evaluation with actual users.

5. ACKNOWLEDGMENTS

The CULTURA (contract no. 269973) and the PROMISE network of excellence (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

6. REFERENCES

[1] Agosti, M., Conlan, O., Ferro, N., Hampson, C., and Munnelly, G. (2013). Interacting with Digital Cultural Heritage Collections via Annotations: The CULTURA Approach. In Marinai, S. and Marriot, K., editors, *Proc. 13th ACM Symposium on Document Engineering* (DocEng 2013), pages 13-22. ACM Press, New York, USA.

[2] Agosti, M., Ferro, N., Frommholz, I., and Thiel, U. (2004). Annotations in Digital Libraries and Collaboratories - Facets, Models and Usage. In Heery, R. and Lyon, L., editors, *Proc. 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*, pages 244-255. Lecture Notes in Computer Science (LNCS) 3232, Springer, Heidelberg, Germany.

[3] Agosti, M., Bonfiglio-Dosio, G., and Ferro, N. (2007). A Historical and Contemporary Study on Annotations to Derive Key Features for Systems Design. *International Journal on Digital Libraries*, 8(1):1-19.

[4] Agosti, M. and Ferro, N. (2008). A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems (TOIS)*, 26(1):3:1-3:57.

[5] Brickley, D. and Guha, R.V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-schema/

[6] Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., and Schuldt, H. (2007). *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*. ISTI-CNR at Gruppo ALI, Pisa, Italy. http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

[7] Carmel, D., Zwerdling, N, and Yogev. S. (2012) Entity Oriented Search and Exploration for Cultural Heritage Collections. *WWW'12, EU Track*

[8] Guy I., Avraham U., Carmel D., Ur S., Jacovi M., Ronen I.. Mining expertise and interests from social media. *In Proceedings of the 22nd international conference on World Wide Web (WWW '13), pp. 515-526.*

[9] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. (2012). The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In *Progress in Cultural Heritage Preservation - 4th International Conference (EuroMed 2012)*, pages 668-675. Lecture Notes in Computer Science (LNCS) 7616, Springer, Heidelberg, Germany.

[10] Hampson, C., Lawless, S., Bailey, E., Yogev, S., Zwerdling, N., Carmel, D., Conlan, O., O'Connor, A. and Wade, V. (2012). CULTURA: A Metadata-Rich Environment to Support the Enhanced Interrogation of Cultural Collections. In *Metadata and Semantics Research - 6th Research Conference (MTSR 2012)*, pages 227-238. Communications in Computer and Information Science (CCIS) 343, Springer, Heidelberg, Germany.

[11] Ferro, N. (2009). Annotation Search: The FAST Way. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 15-26. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany.

[12] OASIS Search Web Services Technical Committee (2012). searchRetrieve: Part 5. CQL: The Contextual Query Language Version 1.0. http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part5-cql.pdf.

[13] Salton, G., Fox, E. A., and Wu, H. (1983). Extended Boolean Information Retrieval. *Communications of the ACM (CACM)*, 26(11):1022–1036.

[14] Yogev, S., Roitman, H., Carmel, D. and Zwerdling, N. Towards Expressive Exploratory Search Over Entity-Relationship Data. *Proceedings of the 21st international conference companion on World Wide Web (WWW) April 16 2012, Lyon, France*