Enriching Digital Cultural Heritage Collections via Annotations: The CULTURA approach*

Maristella Agosti[1], Owen Conlan[2], Nicola Ferro[1], Cormac Hampson[2], Gary Munnelly[2], Chiara Ponchia[1] and Gianmaria Silvello[1]

[1]University of Padua, Italy

{agosti, ferro, silvello}@dei.unipd.it, ponchia.chiara.1@studenti.unipd.it

[2] Trinity College Dublin, Ireland

owen.conlan@scss.tcd.ie, hampsonc@cs.tcd.ie, munnellg@tcd.ie

Abstract. This paper introduces the main characteristics of the CULTURA project and of the IPSA collection, which constitutes one relevant use case presently in use in the CULTURA environment. We describe the innovative annotation features of the CULTURA portal for digital humanities; these features are aimed at improving the interaction of users with digital cultural heritage (CH) content. The annotation functions consist of two modules: the FAST annotation service as back-end and the CAT Web front-end integrated in the CULTURA portal. The annotation service also adds a semantic layer over the managed collections enhancing discoverability and interoperability capabilities of cultural heritage resources. A user-based evaluation of the environment has been carried out.

1. INTRODUCTION

The main aim of the CULTURA project[1] has been to create an innovative environment in which users with a range of different backgrounds and expertise can collaboratively explore, interrogate, interact with, and interpret complex and diverse digital Cultural Heritage (CH) collections. At the conclusion of the project, the resulting environment is a system, which has pushed forward the frontiers of technology in the creation of community and content aware interfaces to digital humanities collections.

The CULTURA environment adopts a service-oriented approach to offer a rich and engaging experience for different user categories, which range from academic and professional users to the general public. The services are conceived and developed to be applicable to a wide variety of cultural collections. The potential generality of the environment is demonstrated by the fact that CULTURA is supporting different use cases; one of those is represented by the IPSA[2] collection – a digital archive of illuminated manuscripts – while the other major archive is the 1641 Depositions, which is a collection of noisy text documents, mainly of a legal nature, dating from the 17th Century.

In both collections, the managed digital objects – either scanned illuminated manuscripts or legal documents – are described by appropriate metadata, according to a traditional record-centric approach. One of the goals of the project was to exploit an improved user engagement and interaction with the managed CH artefacts in order to semantically enrich them with a superimposed layer of user-provided information. This required a move from a traditional record-centric approach to a resource-centric one, opened towards Linked Open Data (LOD) and a better sharing of resources.

[1] http://www.cultura-strep.eu/

[2] http://ipsa.dei.unipd.it/en_GB/home

*An extended version of this paper appears in: M. Agosti, O. Conlan, N. Ferro, C. Hampson, G. Munnelly. 2013. Interacting with digital cultural heritage collections via annotations: the CULTURA approach. In *Proceedings of the 2013 ACM Symposium on Document Engineering* (DocEng '13). ACM, New York, NY, USA, 13-22.

In this paper, we focus on IPSA (*Imaginum Patavinae Scientiae Archivum*), a digital archive of illuminated manuscripts that includes both astrological codices and herbals produced mainly in the Veneto region, in Northern Italy, during the XIV and XV centuries. The digital archive has been created from the corpus of historical and innovative illustrations produced in the centuries where the Paduan School were influential. The online archive was created specifically for professional researchers in the History of Illumination; due to involvement in the CULTURA project, it was decided to open the archive to other categories of users, such as non-domain professional researchers, student communities and the general public, who can all greatly benefit from the semantically enriched resources.

The History of Art provides fertile ground for research into semantically enriched metadata and LOD; indeed, in History of Art the main way to produce new knowledge is to reveal connections between different items (illuminations, pictures, frescos) that can cast new light on an artist, an artistic movement or an art-historical period. The most valuable connections are the unexpected ones linking elements that may seem to have very few features in common. Indeed, in this domain, important discoveries have often been found thanks to associations and connections of items that emerged by chance in the research path identified by domain experts. Therefore, within CULTURA, it was decided that the central tool for allowing researchers in History of Art to discover new knowledge and unveil new links and relationships among cultural heritage resources would be a semantic annotation tool, called FAST-CAT. This software enables semantically-typed links to be superimposed over the managed digital objects, the traditional record-centric metadata, and Web resources in general. Besides being semantically typed, these links can include full-fledged multimedia content, which allows for rich description and explanation of the link and provides added value to both specialist users and the general public.

The paper is structured as follows: Section 2 gives a critical account of the most relevant work and approaches to the design and management of digital annotations. Section 3 illustrates the adopted annotation model together with the main characteristics of the search model. Section 4 introduces the annotation interaction model that has been envisaged and implemented in the environment, and a specific account on the anchoring of annotations is reported. Section 5 reports about the user-based evaluation of the CULTURA environment. Section 6 draws some final remarks.

2. RELATED WORK

As reported in [1, 2], there are different viewpoints about what an annotation is. We can consider annotations to be metadata, content, a form of context, as hypertext, or as dialog acts. In the context of the CULTURA environment three of those viewpoints are more relevant: metadata, content, and hypertext with semantically typed links.

The role of annotations in digital humanities is well known and documented [2, 6]. Subsequently, many different tools, which allow for the annotation of digital humanities content, have been developed. Unfortunately, tools designed specifically for an individual portal are typically only compatible with that system. More general solutions, which can be easily distributed across various sites, have been developed, but these systems often have limited functionalities, e.g. only enabling the annotation of a single content type or not having sharing features [1]. Many different web-centric proposals have been envisaged, including the Annotea project developed by the World Wide Web Consortium (W3C)[3]. Starting from the work done on Annotea, other relevant efforts were developed, in particular the Open Annotation Collaboration[4], which also focussed on humanities [11], and the Annotation Ontology[5]. Those efforts can be considered predecessors of the Open Annotation Community Group[6], which is the active W3C

[3] http://www.w3.org/2001/Annotea/

[4] http://www.openannotation.org/

[5] http://www.jbiomedsem.com/content/2/S2/S4

[6] http://www.w3.org/community/openannotation/

group that has published the Open Annotation Core Data Model[7]. This model specifies an interoperable framework for creating associations between related resources and annotations, using a methodology that conforms to the Architecture of the World Wide Web. This model has the potential to become a standard and to be widely adopted.

The approach that has been chosen in CULTURA is very much that of a service-centric rather than a web-centric environment, but the defined concepts ensure that all of the modelling and architectural requirements are covered similarly to the relevant web-centric efforts that have been mentioned. FAST-CAT (Flexible Annotation Semantic Tool - Content Annotation Tool) is a generic annotation system that was developed as part of the CULTURA project [7, 8] and it directly addresses the challenge of providing a convenient and powerful means of annotating digital content.

3. FAST ANNOTATION MODEL

The FAST annotation service adopts and implements the formal model for annotations proposed by [2], which captures both syntactic and semantic aspects of annotations. According to this model, an annotation is a compound multimedia object, which is constituted by different signs of an 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[8], 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.

Figure 1 shows an example of annotation, which summarizes the discussion so far. The annotation, with identifier `a1`, is authored by the user `ferro`. It annotates an illustration from the *Carrarese Herbal*, f. 162v, whose identifier is `http://ipsa.ipsa-project.org/ipsa-web/r/illustration/135` and which belongs to the IPSA digital archive. The annotation relates to another illustration from the *Roccabonella Herbal*, f. 42r, whose identifier is `http://ipsa.ipsa-project.org/ipsa-web/r/illustration/1213` in the IPSA digital archive; in addition, it relates also to the DBpedia page of the *Carraresi family*, `http://dbpedia.org/resource/Carraresi/`, which endorsed the production of the *Carrarese Herbal*.

In particular, `a1` annotates two distinct parts of the *Carrarese Herbal*. It annotates a region of the illustration representing a cucumber by using a textual sign whose content is "*This illumination presents an extraordinary search for realism*" and whose meaning is a comment in the RDFS namespace, i.e. a comment according to the RDF Schema [4]. It also annotates a region of the manuscript with a textual sign whose content is "*The text is written in an elegant littera textualis*" and whose meaning is also a comment in the RDFS namespace. Note how the content of the sign is plain text in the first case and HTML in the second case 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.

`a1` relates the *Carrarese Herbal* to the *Roccabonella Herbal*, in particular to a region of the illustration representing a cucumber, with a textual sign whose content is "*The Roccabonel-*

[7] http://www.openannotation.org/spec/core/
[8] http://dublincore.org/

la Herbal illumination is clearly copied from the Carrarese Herbal one, as it shows the same disposition of the elements of the plant in the page […]." and whose meaning is to be copied from another illustration in the IPSA namespace. This annotation thus represents the outcomes of the actual work of a historian of art, who conducted his/her own research on these two herbals, to determine that one was copied from the other.

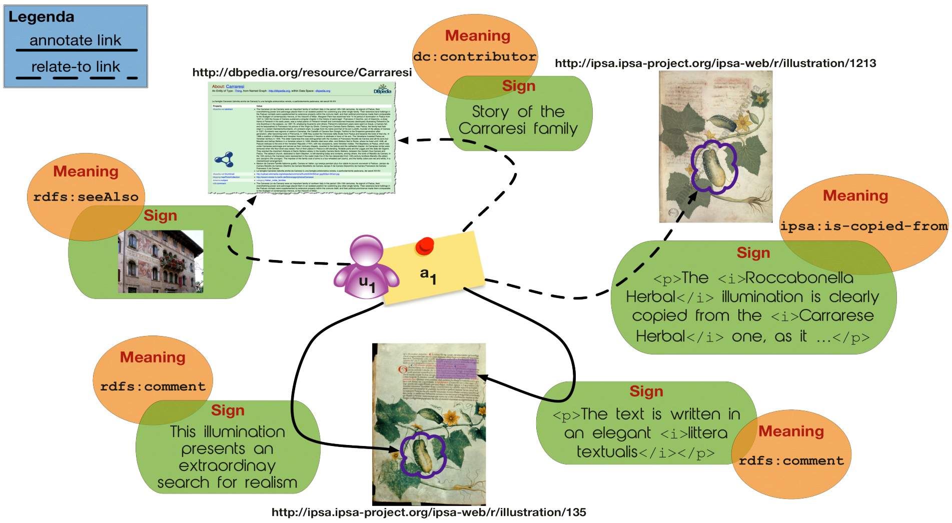

Legenda
annotate link
relate-to link
http://dbpedia.org/resource/Carraresi
Meaning
dc:contributor
Sign
Story of the Carraresi family
http://ipsa.ipsa-project.org/ipsa-web/r/illustration/1213
Meaning
rdfs:seeAlso
Sign
u_1
a_1
Meaning
ipsa:is-copied-from
Sign
<p>The <i>Roccabonella Herbal</i> illumination is clearly copied from the <i>Carrarese Herbal</i> one, as it ...</p>
Meaning
rdfs:comment
Sign
This illumination presents an extraordinay search for realism
Sign
<p>The text is written in an elegant <i>littera textualis</i></p>
Meaning
rdfs:comment
http://ipsa.ipsa-project.org/ipsa-web/r/illustration/135

Figure 1. Example of annotation.

Moreover, `a1` relates the *Carrarese Herbal* to the DBpedia page of the *Carraresi family*, which endorsed the herbal, with two signs: a textual sign whose content is "*Story of the Carraresi family*" and whose meaning is a contributor from the Dublin Core metadata schema; and, an image sign with a picture of a building of the *Carraresi family*, whose meaning is "see also" in the RDFS namespace.

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 structure can be seen as a semantic layer built upon the documents, which serves both for discoverability and interoperability of resources. Indeed, this layer can be straightforwardly exploited to open up annotations and related descriptions as Linked Open Data (LOD) on the Web. For instance, the graph shown in Figure 1 is ready to be shared as LOD. Furthermore, by means of the *relate-to* links annotation `a1` is connected to relevant external LOD resources (i.e. a DBpedia resource). Links towards external resources are the foremost means for augmenting discoverability. The FAST annotation model thus may improve discoverability without requiring to map local resources to RDF or to directly expose them as LOD on the Web. This approach also augments the interoperability potential of the CULTURA environment, because it allows for relating LOD and non-LOD resources via annotations without requiring mapping and synchronization, while respecting visibility (e.g. private or shared) of the resources.

3.1 Search Model

The presence of both structured and unstructured content within the managed resources calls for different types of search functionalities, since structured content can be dealt with *exact match* searches while unstructured content can be dealt with *best match* searches. These two different type of searches may need to be merged together in a query if, for example, the user wants to retrieve annotations by a given author about a given topic. This could be expressed by a boolean AND query which specifies both the author (structured part) and the content (unstructured part) of the annotations to be searched. Nevertheless, boolean searches are best suited for dealing with *exact match* searches and they need to be extended to also deal with *best match* searches. Therefore, any search strategy must be able to express complex conditions that involve both *exact* and *best match* searches. The "P-norm" extended boolean model proposed by [10] is capable of dealing with and mixing both *exact* and *best match* queries, since it is an intermediate between the traditional boolean way of processing queries and the vector space processing model. Indeed, on the one hand, the P-norm model preserves the query structure inherent in the traditional boolean model by distinguishing among different boolean operators (and, or, not); on the other hand, it allows us to retrieve items that would not be retrieved by the traditional boolean model due to its strictness, and to rank them in decreasing order of query-document similarity. Moreover, the P-norm model is able to express queries that range from pure boolean queries to pure vector-space queries, thus offering great flexibility to the user.

The FAST Context Set [5] has been defined in order to provide a uniform query syntax to FAST by using the Contextual Query Language (CQL) [9], developed and maintained by the Library of Congress in the context of the Z39.50 Next Generation (ZING) project[9]. FAST provides conformance to CQL up to Level 2. For example, a possible query to search for information about the Roccabonella herbal and where it is copied from is:

```
           annotation.general = Roccabonella
                    and/match==fuzzy
    annotation.concept.identifier = is-copied-from
```

where the first clause is a *best match* query, the second clause is an *exact match* query and a relaxed boolean search is performed.

4. CAT ANNOTATION INTERACTION MODEL AND TOOL

CAT is a Web annotation tool developed with the goal of being able to annotate multiple types of documents and assist collaboration in the field of digital humanities. At present, CAT allows for the annotation of both text and images. The current granularity for annotation of text is at the level of the letter. For image annotations, the granularity is at the level of the pixel. This allows for extremely precise document annotation, which is very relevant to the Digital Humanities domain due to the variety of different assets that prevail.

There are two types of annotation, which may be created using CAT; a targeted annotation and a note. The former is a comment associated with a specific part of a document. This may be a paragraph, a picture or an individual word – as shown in Figure 2 – but the defining feature is that the text is directly associated with a specific subset of the digital resource.

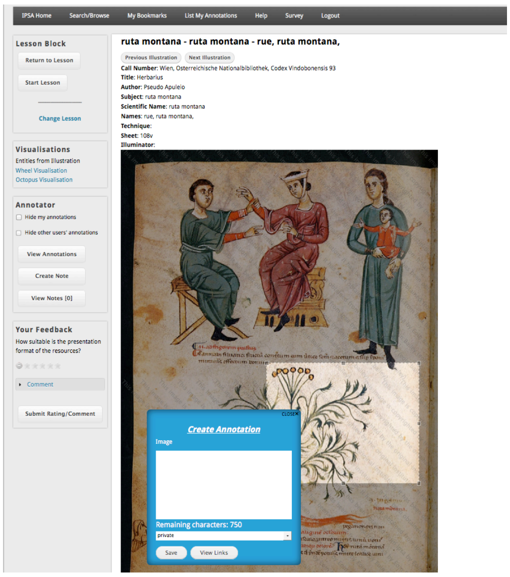
Conversely, a note is simply attached to the document as it is depicted in Figure 3 where several notes have been associated to a document. A note is not associated with a specific aspect of the artifact and typically serves as a general comment about the document as a whole.

In addition to allowing a user to comment on document text, the annotations created using CAT allow an individual to link their annotations to other, external sources. This is beneficial for teachers using digital cultural collections, for students (from primary to university level), as well as experienced researchers. As can be seen in Figure 1, the addition of links to a resource can enrich the amount of information it contains. Each link has comment text associated with it allowing a researcher to explain why this specific link is important or how it supports their argument.

While CAT is beneficial for researchers and educators, it is also being used as an important source of user data for the content provider. For a digital humanities site, annotations can provide an insight into which entities are of interest to a user. If a user is frequently annotating a document, it is likely that this document is of interest to him. Furthermore, if the text

[9] http://www.loc.gov/standards/sru/

being annotated is analysed, it may be possible to discern specific entities of interest within the document. CAT gives access to targeted sections of a document. Simply by selecting a region of interest (within text or images), a toolbar is presented which provides the user with a button to launch the annotation tool. This toolbar is now exposed to other services within the portal, allowing for live interfacing with a document.

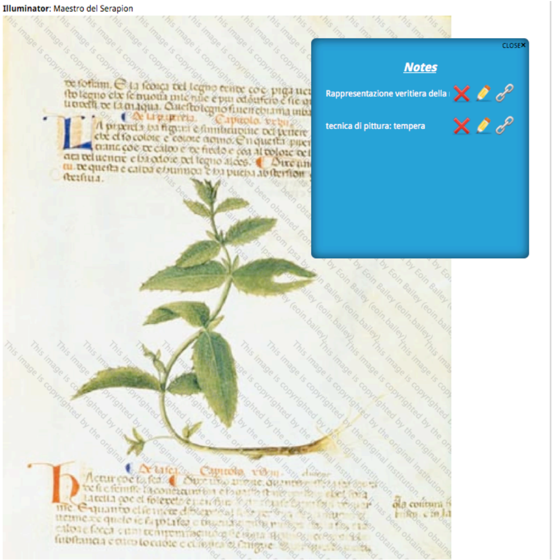


ruta montana - ruta montana - rue, ruta montana,

Call Number: Wien, Osterreichische Nationalbibliothek, Codex Vindobonensis 93
Title: Herbarius
Author: Pseudo Apuleio
Subject: ruta montana
Scientific Name: ruta montana
Names: rue, ruta montana,
Technique:
Sheet: 108v
Illuminator:

Remaining characters: 750
private

Figure 2. User can create an annotation on an image of interest.



Illuminator: Maestro del Serapion

Rappresentazione veritiera della
tecnica di pittura: tempera

Figure 3. Some notes have been associated to a document.

5. USER-BASED EVALUATION OF FAST-CAT

Both the FAST annotation model and the CAT model and tool (FAST-CAT) have been applied to the CULTURA environment and they provided adaptive and personalized access to the IPSA historical collection. FAST-CAT has been integrated into the environment in order to provide users with an additional means of interacting with the portal, as well as for providing feedback on CULTURA's user model that stores a user's interests. To investigate the quality of CULTURA technologies in terms of the opportunities, experiences and added value they provide to users, and to enable an iterative refinement of the research environment, evaluation work was an inherent part of the project work. Evaluations were carried out during all three years of the project, with evaluations in phase 1 (baseline evaluation), 2 (mid-term) and 3 (final one). FAST-CAT was evaluated as soon as was made available and in the final phase. In total 12 user trials in the context of the IPSA collection, and spread over the whole spectrum of user groups were conducted.

FAST-CAT proved popular with users of the initial system and was extended to support private annotations, public annotations and group annotations [12]. This enables users with different needs (educators creating annotations for their class; students collaborating on a project; historians documenting their private insights etc.) to have their annotations shared at an appropriate level.

Regarding the assessment of different CULTURA services and functionalities, the following qualities were of main interest: adaptation quality, visualisation quality, collaboration support, research support given by the annotation tool, the bookmarking functionality, text normalisation, and narrative lessons. Results on the research support given by the annotation tool were good, as were the results on collaboration support through the sharing annotations. The usability of the annotation tool was also rated high[10].

6. CONCLUSIONS AND FUTURE WORK

It is the belief of the authors that FAST-CAT has huge potential as an annotation tool within the digital humanities field. Indeed, it demonstrated the feasibility of transitioning from a traditional digital archive with a record-centric approach, towards a resource-centric one with semantically enriched information provided by actively engaging users via digital annotations. However, it is still a young tool with much room for future expansion and enhancement. Work has already started to further push the boundaries of IPSA by considering the group of researchers that interact with the collection via FAST as entities modelled by the system [3]. These entities can then be exploited by creating new services. For instance, we offer a way to construct a social network of researchers based on inferred relationships. An inferred relationship between researchers could be derived from their co-annotation of a certain illuminated manuscript. The social network between researchers can serve as the base for social services, such as personalized content recommendation and adaptive personalized search [3].

The process of discovering new unexpected connections among cultural heritage artefacts – i.e. a process that can be defined as 'serendipity' – enabled by the FAST annotation model, is especially encouraged by the LOD paradigm where meaningful links between entities allow us to move across diverse and apparently unrelated knowledge domains. History of Art is a particularly well-suited domain where algorithms fostering serendipity within LOD can be designed, developed and evaluated because it provides: rich and heterogeneous information needs, ample structured and unstructured resource datasets and a proactive community accustomed to seeking new semantic connections between entities. Future work will focus on these aspects and

[10] The results of the evaluation trials are reported in a restricted deliverable, but a new evaluation effort is in progress and preliminary results will be presented at the conference.

will further extend the functions and services provided to the users of the CULTURA environment.

ACKNOWLEDGMENTS

The CULTURA (contract no. 269973) project, as part of the 7th Framework Program of the European Commission, has partially supported the reported work.

7. REFERENCES

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

[2] 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.

[3] Agosti, M., Ferro, N., Porat, S., and Rabinovich, E. (2014). Enhancing Digital Cultural Heritage Collections with Social Network Capabilities. In *Proc. of the PATCH workshop: The Future of Experiencing Cultural Heritage,* Part of the IUI 2014 Conf. Haifa, Israel.

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

[5] Ferro, N. (2009). Annotation Search: The FAST Way. In *Proc. 13th European Conf. on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, LNCS 5714, Springer, pp 15-26

[6] Ferro, N. and Silvello, G. (2013). NESTOR: A Formal Model for Digital Archives. *Information Processing & Management*, 49(6):1206-1240.

[7] 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 Int. Conf. (EuroMed 2012)*, LNCS 7616, Springer, pp 668-675.

[8] 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 Conf. (MTSR 2012)*, CCIS 343, Springer, pp 227-238.

[9] 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.

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

[11] Sanderson, R. and Van de Sompel, H. (2010). Making web annotations persistent over time. In *Proc. of the Joint Int. Conf. on Digital Libraries (JCDL 2010)*. ACM, pp 1-10.

[12] Sweetnam, M., Agosti, M., Orio, N., Ponchia, C., Steiner, C., Hillemann, E., Ó Siochrú, M., and Lawless, L. (2012). User Needs for Enhanced Engagement with Cultural Heritage Collections. In *Proc. of the 2nd Int. Conf. on Theory and Practice of Digital Libraries (TPDL 2012)*, Paphos, Cyprus, LNCS Vol. 7489, Springer, 2012, pp 64-75.