Interacting with Digital Cultural Heritage Collections via Annotations: The CULTURA Approach

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

Owen Conlan
Knowledge and Data
Engineering Group
Trinity College
Dublin, Ireland
Owen.Conlan@scss.tcd.ie

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

Cormac Hampson
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
hampsonc@cs.tcd.ie

Gary Munnelly
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
munnellg@tcd.ie

ABSTRACT

This paper introduces the main characteristics of the digital cultural collections that constitute the use cases presently in use in the CULTURA environment. A section on related work follows giving an account on efforts on the management of digital annotations that are pertinent and that have been considered. Afterwards the innovative annotation features of the CULTURA portal for digital humanities are described; those features are aimed at improving the interaction of non-specialist users and general public with digital cultural heritage 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 features have been, and are being, tested with different types of users and useful feedback is being collated, with the overall aim of generalising the approach to diverse document collections and not only the area of cultural heritage.

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.*

Keywords

Cultural heritage collections, digital cultural heritage collections, digital libraries and archives, digital library system, annotation, hypertext, digital humanities, adaptive environment.

DocEng'13, September 10–13, 2013, Florence, Italy.
Copyright © 2013 ACM 978-1-4503-1789-4/13/09…$15.00.
http://dx.doi.org/10.1145/2494266.2494288

1. INTRODUCTION

The CULTURA project[1] aims to create an innovative information and communications technology (ICT) environment in which users with a range of different backgrounds and expertise can collaboratively explore, interrogate and interpret complex and diverse digital cultural heritage collections. At the conclusion of the project, the resulting environment will be 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 is service oriented and is composed of a set of services which are integrated to create a rich and engaging experience that supports users of different 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 that are represented by the IPSA[2] and 1641[3] collections, which differ in morphology, language, modality and metadata. This means that the environment and the supported services need to consider the peculiarities of different documents and different ways of making use of them by diverse categories of users. One of the supported services which must be designed and made available, taking into specific account the peculiarities of the documents of different collections, is the annotation service.

The paper is structured as follows: Section 2 presents the two use cases that at present are managed by the environment, and Section 3 gives a critical account of the most relevant work and approaches to the design and management of digital annotations. Section 4 illustrates the adopted annotation model together with the main characteristics of the search model. Section 5 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 6 details the

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

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

[3] http://1641.tcd.ie/about.php

architecture that manages the interaction with annotations, and section 7 reports on the characteristics of the CULTURA environment and on the efforts that have been put in place for evaluating its annotation features. Section 8 presents some initial results of a user trial that was performed with FAST-CAT. Finally, Section 9 concludes the paper and presents some insights on future developments.

2. USE CASES

The CULTURA environment is supporting different use cases that are represented by the IPSA and 1641 collections, which differ in many characteristics that are presented briefly in this section.

IPSA (*Imaginum Patavinae Scientiae Archivum* - Archive of images to support the study of scientific research at Padua University) is 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 conceived starting from the corpus of the historical and very innovative illustrations produced in the centuries under the influence of the Paduan School. The online archive was created specifically for professional researchers in History of Illumination to allow them to compare the illuminated images held in the collection and verify the development of a new realistic way of painting closely associated with the new scientific studies that were flourishing at the University of Padua in the XIV century, particularly thanks to the teaching of Pietro d'Abano. Disclosing new relationships between images is one of the main purposes of research in art history, because it brings further knowledge on a painter or an illuminator, on a work of art, or on a whole specific artistic period. According to this particular user requirement, in IPSA professional researchers are provided with tools that allow them to link and annotate images, so they are able to keep track of their considerations on the illuminations and their relations [5].

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. Between May and October 2012 relevant data from IPSA was selected to constitute the collection to be imported in the CULTURA environment for use as a case study to test the new environment and its functions.

The 1641 Depositions are 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 or 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.

From a technical perspective, the 1641 Depositions represent a textually rich digital humanities collection. This is in contrast to the IPSA collection, which is highly visual in nature and presents different challenges for digital humanists. The Depositions have active communities of interest because of their wider social and historical implications that transcend geographical and chronological boundaries and continue to shape opinions and values to this day. The depositions display important similarities to much of the user-generated content found on the World Wide Web today. They are inconsistent in almost every aspect, including spelling, punctuation, case and language. The entire collection of 1641 Depositions have now been processed and integrated into the CULTURA environment. This processing involved the normalisation of text, the extraction of entities (people, places, dates etc.) and their relationships, as well as the use of social network analysis. Once processed and integrated into CULTURA, a range of services to help the exploration and analysis of the 1641 Depositions (and collaboration around the collection) are made available to users.

3. RELATED WORK

Almost everybody is familiar with annotations and has his or her own intuitive idea about what they are. These are drawn from personal experience and the habit of dealing with types of annotation in everyday life, e.g. jottings for the shopping, taking notes during a lecture or adding a commentary to a text. This intuitiveness makes annotations especially appealing for both researchers and 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.

Many user studies have been conducted to understand annotation practices and to discover common annotation patterns. Marshall [26] has categorised annotations along several dimensions that reflect the form annotations may take on. Others have focussed on the design and development of document models and systems which support annotations in specific classes of management systems, such as digital libraries, the Web, laboratory systems and working groups, databases, and adaptive environments. This work has led to different viewpoints about what an annotation is; as reported in [7]. 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. Taking this into account we briefly and critically examine what these three points mean, and we make reference to previous work.

Annotations are metadata because they can be considered as additional information which concerns existing content, i.e. they are metadata, as they clarify in some way the properties and the semantics of the annotated content. An example of the use of annotations as metadata is MPEG-7 [27, 28], which is an ISO/IEC standard developed by the MPEG (Moving Picture Experts Group) committee. MPEG-7 serves for annotating and describing multimedia content data, and supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible.

Another example, in the context of the database management area, sees annotations as "information about data such as provenance, comments, or other types of metadata" [12], which is a sort of data that is added to an existing database. It could be additional data that for whatever reason cannot be stored in the original database, or it could be some form of metadata such as comments, probabilities, timestamps that are not normally regarded part of the basic database design [24].

It has recently been observed that, in order to determine how annotations should be propagated through database queries, it is necessary to have some structure on them. Although various forms of annotation have been considered in some detail, each form has been considered in isolation; Buneman and Kostylev have proposed a hierarchical model of annotation in which there is no absolute distinction between annotation and data [14], in this case annotations can be considered both as metadata and content as in [29] where annotations are additional content which concern existing content, and they increase existing content by providing an additional layer of content that elucidates and explains the existing one. This viewpoint about annotations entails an intrinsic dualism between annotation as content enrichment and annotation as stand-alone document [2]: 1) annotations as content enrichment are considered as mere additional content regarding an existing document and as a result they are not autonomous entities but in fact they rely on previously existing information resources to justify their existence; 2) annotations as stand-alone documents are considered as real documents and are autonomous entities that maintain some sort of connection with an existing document.

Annotations can allow the creation of new relationships between existing content, by means of links that connect annotations together with existing content. Using this viewpoint we can consider that existing content and annotations constitute a hypertext [4, 6], according to the definition of hypertext provided in [1]. This hypertext can be exploited to provide alternative navigation and browsing capabilities, but also to offer advanced search functionalities. Furthermore, [26] considers annotations as a natural way of creating and increasing hypertexts that connect information resources in a digital library management system by actively engaging users. The hypertext which exists between information resources and annotations enables different annotation configurations that are threads of annotations, i.e., an annotation made in response to another annotation, or sets of annotations, i.e. a bundle of annotations on the same information resource [2, 3].

As has been pointed out, 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 role of annotations in digital humanities is well known and documented [3, 7, 9-11, 18]. The authors of [38] propose a general framework for annotating large archives of historical image manuscripts. The work is similar in spirit to the work that has been done for the IPSA application on the automatic discovery of relationships among images in illuminated manuscripts [8]; however in [38] the authors are focusing on the lower level primitives to support such work using different feature spaces such as shape, colour and texture, and their relevant contribution is in introducing a novel technique for calculating weighting parameters, without having a labelled training data collection.

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 functionality e.g. only enabling the annotation of a single content type or not having sharing features [31, 34].

Many different web-centric proposals have been envisaged, including the project developed by the World Wide Web Consortium (W3C) activity on Annotea[4] [23]. Starting from the work done on Annotea, other relevant efforts have been developed, in particular the Open Annotation Collaboration[5], that also focussed on humanities [34], and the Annotation Ontology[6]. Those efforts can be considered predecessors of the Open Annotation Community Group[7], which is the active W3C group that has published the Open Annotation Core Data Model[8]. 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. Before designing the annotation tools for the CULTURA project, the Text-Image Linking Environment (TILE) [35] was assessed. However TILE, was not adopted, because it is a web-based tool for creating and editing image-based electronic editions and digital archives of humanities texts, and the creation of electronic editions is not required in the context of CULTURA.

FAST-CAT (Flexible Annotation Semantic Tool - Content Annotation Tool) is a generic annotation system that is being developed as part of the CULTURA project [21-22] and that directly addresses the challenge of providing a convenient and powerful means of annotating digital content. The remainder of this paper reports on the characteristics of FAST, the backend service providing powerful annotation functionalities, and CAT, the frontend Web annotation tool, and discusses how its features are tackling important challenges within the Digital Humanities field. A key aspect of CULTURA is the development of an online environment that empowers users at various levels of expertise to investigate, comprehend and contribute to digital cultural collections. FAST-CAT is a key component of this environment and is currently being trialled with the help of different user groups.

4. FAST ANNOTATION MODEL

The FAST annotation service adopts and implements the formal model for annotations proposed by [9], which captures both syntactic and semantic aspects of annotations. The adopted model has been also embedded in the reference model for digital libraries developed by DELOS, the European network of excellence on digital libraries [15].

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

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

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

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

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

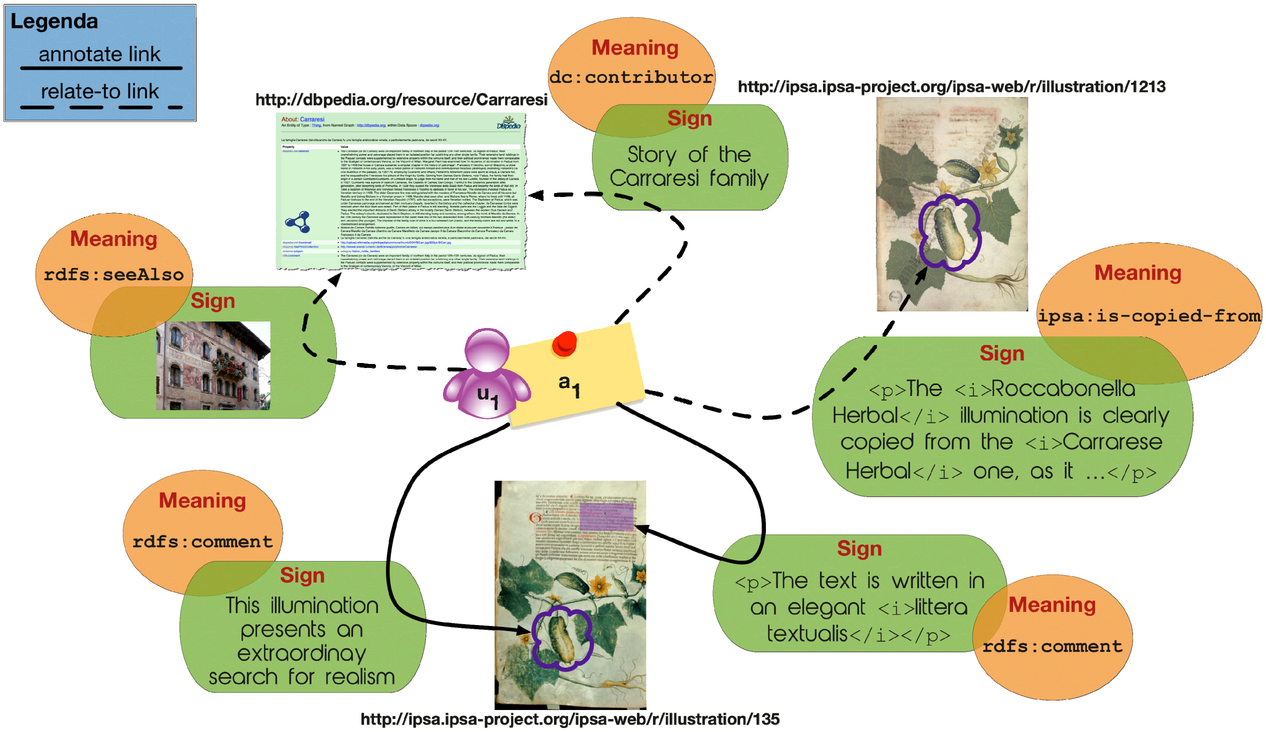

Legenda
annotate link
relate-to link
http://dbpedia.org/resource/Carraresi
Meaning
dc:contributor
Sign
Story of the Carraresi family
http://ipsa.ipsa-project.org/ipsa-web/r/illustration/1213
Meaning
rdfs:seeAlso
Sign
u1
a1
Meaning
ipsa:is-copied-from
Sign
<p>The <i>Roccabonella Herbal</i> illumination 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.

According to this model, an annotation is a compound multimedia object which 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[9], 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 to be a comment in the RDFS namespace, i.e. a comment according to the RDF Schema W3C recommendation [13]. 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 to be 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 an illustration representing a cucumber as well, with a textual sign whose content is "The *Roccabonella 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, the same search for realism and the same attention to the light effects on the surface of the leaves, the fruits and the flowers." 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 an historian of art, who conducted his/her own research on these two herbals, to determine that one was copied from the other.

[9] http://dublincore.org/

Moreover, a1 relates the *Carrarese Herbal* to the DPpedia 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 contributor in 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 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.

4.1 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 types 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 somewhat extended to also deal with best match searches. Therefore, we need to envision a search strategy able to express complex conditions that involve both exact and best match searches. The "P-norm" extended boolean model proposed by [33] 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 hypertext that connects documents to annotations calls for a search strategy that takes it into consideration and allows us to modify the score of annotations and/or documents according to the paths in the hypertext. For example, we could consider that an annotation, retrieved in response to a user query, is more relevant if it is part of a thread where other annotations have also been retrieved in response to the same query rather than if it is part of a thread where it is the only annotation that matches the query.

The FAST Context Set [17] has been defined in order to provide a uniform query syntax to FAST by using the Contextual Query Language (CQL) [30], developed and maintained by the Library of Congress in the context of the Z39.50 Next Generation (ZING) project[10]. FAST provides conformance to CQL up to Level 2.

For example, a possible query to search information about the Roccabonella herbal and where it is copied from is:

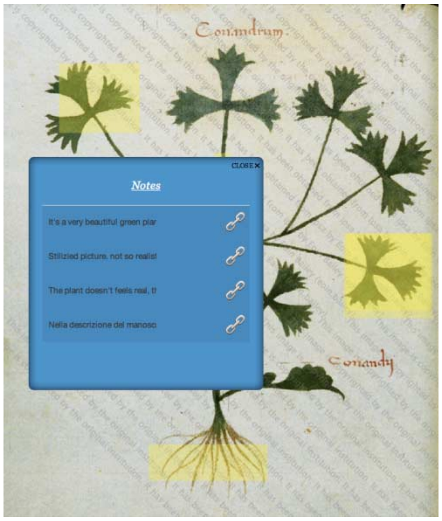
```
           annotation.general = Roccabonella
                   and/match==fuzzy
    annotation.concept.identifier = is-copied-from
```

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

5. CAT ANNOTATION INTERACTION MODEL

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.

CAT can create two types of annotation, the first of which is a targeted annotation – a comment which is associated with a specific part of a document. This may be a paragraph, a picture or an individual word, but the defining feature is that the text is directly associated with a specific subset of the digital resource.



It's a very beautiful green plar

Stilizied picture. not so realist

The plant doesn't feels real, tl

Nella descrizione del manosc

Figure 2. Some notes have been associated to a document.

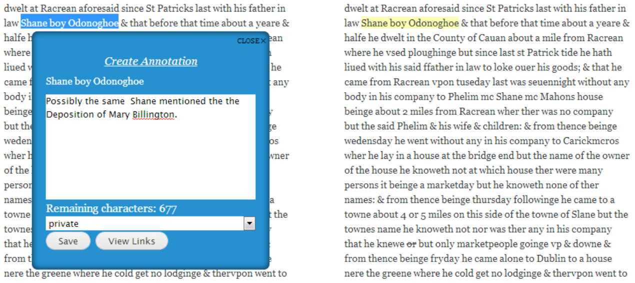
The second type, a note, is simply attached to the document as shown in Figure 2 where some notes have been associated with a resource. A note is not associated with a specific item. Rather, it serves as a general comment about the document as a whole.

The annotations created using CAT allow an individual to link their comments to other, external sources. This is hugely beneficial for teachers using digital cultural collections and 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 greatly 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 point.

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

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

frequently annotating a document, it is likely that this document is of interest to them. Furthermore, if the text being annotated is analysed, it may be possible to discern specific entities of interest within the document. This can be used to drive a personalised recommender service, populating it with entities pertaining to an individual user. A digital humanities site which could recommend resources that are relevant to users would be profoundly useful, and would help improve the effectiveness with which researchers interact with their domain.



Shane boy Odonoghoe

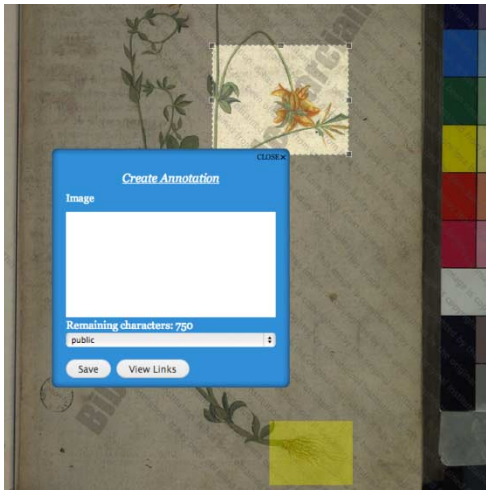
Possibly the same Shane mentioned the the Deposition of Mary Billington.

Remaining characters: 677

private

dwelt at Racrean aforesaid since St Patricks last with his father in law Shane boy Odonoghoe & that before that time about a yeare & halfe he dwelt in the County of Cauan about a mile from Racrean where he vsed ploughinge but since last st Patrick tide he hath liued with his said ffather in law to loke ouer his goods; & that he came from Racrean vpon tuseday last was seuennight without any body in his company to Phelim mc Shane mc Mahons house beinge about 2 miles from Racrean wher ther was no company but the said Phelim & his wife & children: & from thence beinge wedensday he went without any in his company to Carickmcros wher he lay in a house at the bridge end but the name of the owner of the house he knoweth not at which house ther were many persons it beinge a marketday but he knoweth none of ther names: & from thence beinge thursday followinge he came to a towne about 4 or 5 miles on this side of the towne of Slane but the townes name he knoweth not nor was ther any in his company that he knewe ~~or~~ but only marketpeople goinge vp & downe & from thence beinge fryday he came alone to Dublin to a house nere the greene where he cold get no lodginge & thervpon went to

Figure 3. User creates a targeted annotation on a body of text about a person of interest.



Remaining characters: 750

public

Figure 4. User can create an annotation on an image of interest.

5.1 Annotation Pointers

Here we present the means by which CAT identifies the placement of an annotation within a web document using a serialised pointer to a location.

For text, this serialized representation takes the form:

<PathStart>;<OffsetStart>;<PathEnd>;<OffsetEnd>

Where:

- <PathStart> is the path to the element which contains the start of the user's selection.
- <OffsetStart> is the offset into the start element where the beginning of the selected text may be found.
- <PathEnd> is the path to the element which contains the end of the user's selection.
- <OffsetEnd> is the offset into the end element where the end of the selected text may be found.

For images, the form is:

<Path>;<OffsetX>;<OffsetY>;<AnnotationH>;<AnnotationW>

Where:

- <Path> is the path to the annotated image.
- <OffsetX> and <OffsetY> are the position of the upper left corner of the annotation.
- <AnnotationH> and <AnnotationW> are the height and width of the annotation within the image.

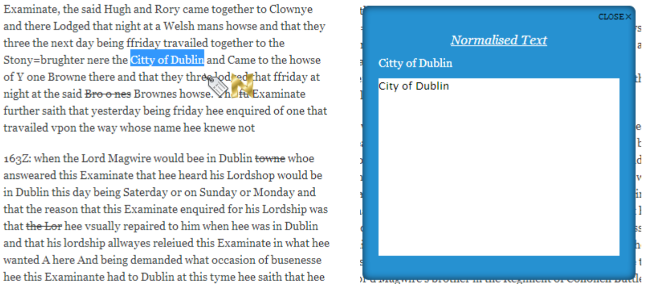
In both cases, the path is computed using a modified version of the open source Okfn annotator [31] range class. In order to improve cross browser compatibility, CAT replaces Okfn's XPath pointers with CSS selectors. There are two reasons for this change. Firstly, different browsers will render pages in different ways, which means that XPath is not always a reliable means of locating a specific element in the markup. Secondly, support for XPath has been removed from current releases of jQuery. CSS selectors, however, are still supported and hence are the more suitable choice.

Additionally, rather than using browser ranges, CAT uses Rangy [32] ranges. Rangy is an open source JavaScript library which creates a virtual representation of a selected range that is independent of the browser being used. Rangy can then map this virtual range to the current page, taking into consideration the browser being used. Pointers are generated with respect to this virtual range so that the result should always evaluate to the same document location regardless of the environment. FAST is flexible enough that it can store this annotation representation without any modification, either to itself or to CAT.

5.2 Expanding Functionality of CAT

Within the context of CULTURA, 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.

By way of example, CULTURA provides a normalization service which resolves anomalies in the archaic text of the 1641 Depositions [25]. Access to this service is now provided to the user via CAT. A new button (based on a default CAT button class) has been added to the toolbar which launches CAT with the normalisation service (See Figure 5). By using this toolbar, the normalization service can access the robust ranges used in FAST-CAT to perform targeted processing of text. While the primary design and focus of CAT was to provide a reliable means of annotating document text, the way it has been implemented means that site developers can easily access its functionalities, thus providing a powerful method of connecting their services with specific parts of a document.



Examinate, the said Hugh and Rory came together to Clownye and there Lodged that night at a Welsh mans howse and that they three the next day being ffriday travailed together to the Stony=brughter nere the Citty of Dublin and Came to the howse of Y one Browne there and that they three lodged that ffriday at night at the said ~~Bro o nes~~ Brownes howse. [illegible] Examinate further saith that yesterday being friday hee enquired of one that travailed vpon the way whose name hee knewe not

163Z: when the Lord Magwire would bee in Dublin ~~towne~~ whoe answeared this Examinate that hee heard his Lordshop would be in Dublin this day being Saterday or on Sunday or Monday and that the reason that this Examinate enquired for his Lordship was that ~~the Lor~~ hee vsually repaired to him when hee was in Dublin and that his lordship allwayes releiued this Examinate in what hee wanted A here And being demanded what occasion of busenesse hee this Examinante had to Dublin at this tyme hee saith that hee

CLOSE ×

Normalised Text

Citty of Dublin

City of Dublin

Figure 5. Normalization exposed to user via CAT

6. ARCHITECTURE

6.1 FAST Architecture

The FAST annotation service adopts a three-layer architecture, separating the data and service logic from the application and interface logic. All the resources managed by the system, e.g. annotations, signs, meanings, users, and so on, are exposed via a REST interface at HTTP level [19] which offers basic operations to create, read, update, delete, and link these resources together. The REST interface also supports the search and retrieval of resources according to the search model and query language described in Section 4.1.

All the resources are exposed in two formats: XML, according to the FAST XML Schema[11], to make resources available in an application neutral way and to favour interoperability; and, JSON to facilitate the design and development of rich and interactive Web 2.0 applications that utilise an AJAX-based approach.

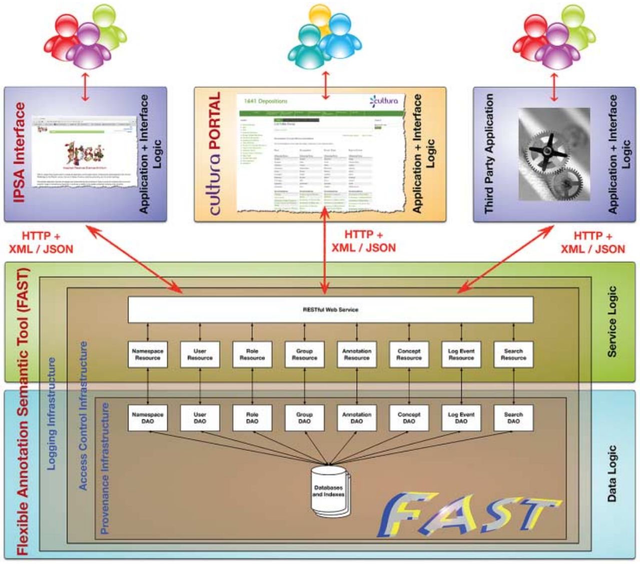


Figure 6. Architecture of the FAST annotation service.

Figure 6 shows the architecture of the FAST annotation service. The Content Annotation Tool (CAT) application and the CULTURA portal represent two possible applications built on top of FAST.

In the data logic layer, a Data Access Object (DAO)[12] takes care of mapping the object oriented representation of the resource to the underlying relational representation of them. The data logic also manages the Provenance Infrastructure, which keeps the history of all the modifications and events related to each resource in order to be able to reconstruct its full history. Each action on a resource originates a provenance event in the form of a statement `<when> <who> <predicate> <what> <why>`, which details when and who performed the given predicate e.g. CREATE, UPDATE etc. on the specified resource (what) for what reason (why) and keeps a dump of the resource itself.

As discussed above, the service logic layer translates from the object oriented representation of the resources to their XML or JSON representations and makes them available via a REST API. So, for example, it is possible to read an annotation performing a HTTP GET to the `/annotation/{id};{ns}` URI, where `id` is the identifier of the annotation, `ns` is its namespace, and in the HTTP header `accept,` you can specify whether XML or JSON has to be returned.

Transferral to the data and service logic layers are the responsibility of the Logging and Access Control infrastructures. The former records all the events happening at the system and database level. Moreover, the interaction at the REST HTTP level is recorded according to the W3C Extended Log File Format [20]. Note that log events are exposed as resources as well, so it is possible to create applications to read, visualise, and search them according to the search model described in Section 4.1. The Access Control Infrastructure takes care of authentication and authorization in a twofold way: user roles define which actions e.g. read annotation or create annotation, a user is allowed to perform; and resource scope and user groups define which resources a user can actually access. For example, a user may be entitled to read annotations by its role but he/she cannot read a specific annotation because of insufficient access permissions.

The FAST annotation service has been developed by using the Java programming language, which ensures good portability of the system across different platforms. We used the PostgreSQL DataBase Management System (DBMS)[13] for the actual persistence of annotations and its full text extension for indexing and searching the full text components of the managed resources. The Apache Tomcat Web container and the Restlet framework have been used for developing the FAST RESTful Web Application.

6.2 CAT Architecture

The architecture of the CAT annotation tool is comprised of two layers; a client-side front end, coded using JavaScript and jQuery, and a Drupal 7 module back end[14], written in PHP, as illustrated in Figure 7.

The front end runs in the user's browser and provides them with a user interface through which they can interact with annotations. When a user has chosen a particular course of action the data is passed into the logic module where their request can be processed. Depending on the nature of the request, certain third party libraries may be used in the procedure. For example, in the

[11] http://ims.dei.unipd.it/data/xml/fast.3.10.xsd

[12] http://www.oracle.com/technetwork/java/dataaccessobject-138824.html

[13] http://www.postgresql.org/

[14] http://drupal.org/

process of annotating a text object, the location of the text in the document must be recorded in a cross platform manner. In order to do this, a representation of the highlighted range is generated using rangy. This is a purely virtual range which means it is slightly slower than using the browser's range, but it has the advantage of being cross platform. Using a modified version of the Okfn path finder, the logic then computes a serialized path to the selected location represented by rangy which can be stored as a pointer in FAST. When annotating images, the process is the same except that jCrop [16] provides details of the selected region rather than Rangy. Retrieving an annotated region is simply the reverse of this process.

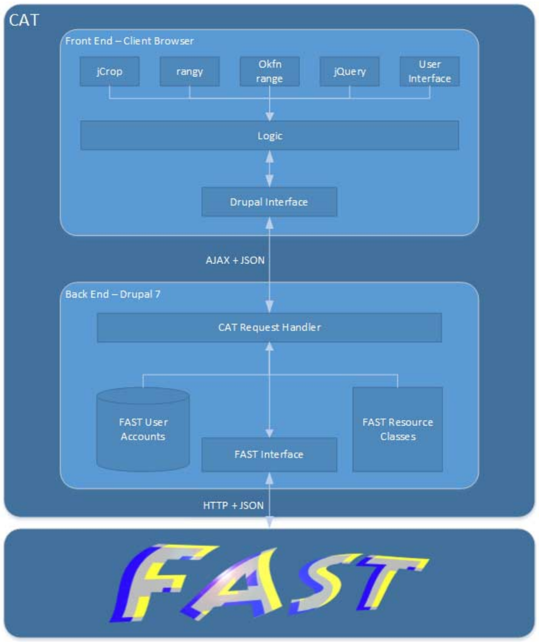


Figure 7. Architecture of the CAT annotation tool.

The representation of an annotation created here is a simplified version of the FAST description of the annotation. This is to minimize the amount of data that a user must send and receive to and from the server. For example, details such as namespaces are added on the back end rather than on the front end (and thus are managed by the site administrator). Furthermore, when managing details such as groups, the user's permissions are derived from the verbose annotation description on the server and then passed as a single value in the simplified representation.

The Drupal 7 module on the back end acts as a relay between FAST and the user. Requests for annotation creation, deletion, download etc. are passed from the front end to a request handler function on the back end. This callback function structures the data sent by the front end so that it conforms to the FAST schema and then generates the HTTP packets to be transferred. There is some logic applied at this point to determine which packets need to be sent and in what order for the request to be fulfilled. Once the system is ready, the packets are sent on to FAST. The Drupal module then waits for a response from the remote service. When one is received, the result is returned to the front end via the same callback function through which the request was initially made.

The choice of a Drupal module as a means of implementation means that adding FAST-CAT to any site using the Drupal CMS should be a very simple process. Additionally, as the Drupal module is only acting as a relay, it should be a relatively simple process to swap out the back end for a more server agnostic implementation, allowing FAST-CAT to be deployed on any website, rather than only those using the Drupal 7 content management system.

Certain requests such as creating and viewing annotations require user authentication by FAST. As FAST is a stand-alone service, it maintains its own record of user accounts and login details. This means that for each user who is registered on the CULTURA site, a separate account must be created for them in FAST. CAT performs this registration automatically.

7. THE CULTURA ENVIRONMENT

CULTURA is a three year, FP7 funded project, scheduled to finish in February 2014. Its main objective is to pioneer the development of personalised information retrieval and presentation, contextual adaptivity and social analysis in a digital humanities context. In its current form, it aims to provide adaptive and personalized access to two historical collections – the 1641 Depositions [36] and IPSA [37].

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 to provide some feedback for CULTURA's user model regarding a user's interests. At present, CULTURA (and by extension FAST-CAT) is being evaluated by three groups of users. A team of MPhil students and professional researchers from Trinity College Dublin are using FAST-CAT as part of their teaching, collaboration and research into the 1641 Depositions. These users will be testing the annotation tool in a free form manner. How they choose to annotate and what content they label is entirely determined by their own needs.

Providing an alternative insight to FAST-CAT is a group of secondary school students from Lancaster who used the annotations as part of a project they were given during a lesson. Their experience was more guided than that of the MPhil students as they were directed to highlight information or points of interest using FAST-CAT and then deliver a presentation using annotations to help with organization. The focus of this lesson was on the 1641 Depositions.

Masters students in Padua will test the image annotation functionality of FAST-CAT as part of their research into the Imaginum Patavinae Scientiae Archivum (IPSA) [37] collections of illuminated manuscripts. Similarly to the MPhil students, the approach of these Masters students to annotating documents will be determined by their own research methodology. The intention is not to guide the users on how to use FAST-CAT, but rather to make them aware of the functionality provided and observe how they choose to apply it.

The various features offered by FAST-CAT and its user interface will be evaluated in detail and comparisons will be drawn between the manner in which different user groups availed of annotations depending on their level of expertise and the type of documents examined. Furthermore, FAST-CAT will also help to drive CULTURA's comprehensive user model by providing the site with updates on the user's behaviour regarding document annotation.

8. EVALUATION

At present, only the results of the trial with 21 secondary school students from Lancaster, mentioned in Section 7, have been collated and analysed. For this evaluation, the students were divided into three groups and provided with activity sheets describing research questions for the 1641 Depositions which they were to investigate using the CULTURA portal.

The annotation tool was provided to give team members a means by which they could draw their peers’ attention to specific areas of the document, share ideas and store personal notes. In this capacity, FAST-CAT excelled, and the tool was used extensively by all parties. When completing a survey at the end of the study, several participants made reference to annotations as a fundamental and valuable tool for this form of collaboration. Indeed 11 of the 21 participants cited the annotation tool as the most helpful service provided by the CULTURA environment.

Despite the positive feedback received about the annotation tool, a number of issues were raised by the students. For instance, there was a difficulty in distinguishing between different users' annotations. All annotations appeared on the page as a yellow highlight over the text. As the number of annotations on a page increased, it became increasingly difficult to identify the owner of a particular comment. While functionality was provided to hide annotations that didn't belong to the individual user, it was clear that better controls were required to help with this aspect of the tool. In the questionnaire that followed this experiment, a number of students mentioned that access to a palette of colours for their annotations, so that each member of a team could produce different coloured highlights on the text, would be desirable. This suggested functionality will be applied to CAT in the next phase of implementation.

As these students were not professional historians, many struggled with the archaic and non-standard version of English that the 1641 Depositions are written in. As such, it was mentioned frequently that an ability to read a normalised version of the text was desirable. As mentioned in Section 5.2 this feature has now been added to CAT, with the ability for users to highlight any text and have it rendered to them in a normalised fashion.

Overall, FAST-CAT was very well received by the test group and will go through several more iterations in the course of the CULTURA project. Evaluation studies are ongoing, and feedback from the entire spectrum of users (professional historians to members of the general public) will be accounted for in the design of the annotation tool.

9. 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. However, it is still a young tool with much room for future expansion and enhancement. Some of the required additions are already known and are currently being developed within the timescale of the project. Others will be dependent on user feedback from test groups as they identify issues they experience within their domains. Some project researchers are working closely with the users to evaluate the usefulness of the environment and to guide its further development/refinement. This process is crucial to the effective design of an environment that will be useful for a range of users who come to the resource, and the collections that it makes available.

As was previously mentioned, it is possible to make FAST-CAT more server agnostic by swapping out the Drupal 7 back end for a more general php script. It is expected that this script will be developed and provided with future versions of FAST-CAT so as to increase the range of portals to which it may be applied. Further to this, another part of the future development of FAST-CAT will be focused on improving the user’s experience. It is intended that the tool be as intuitive and easy to use as possible. How this will be achieved is to be this based on the feedback given by the user groups during the CULTURA trials.

10. ACKNOWLEDGMENTS

The authors would like to thank Chiara Ponchia for her valuable advice on aspects related to the illuminated manuscripts.

We are most grateful to our referees for their very helpful comments.

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.

11. REFERENCES

[1] Agosti, M. (1996). An overview of hypertext. In Agosti, M. and Smeaton, A., editors, *Information Retrieval and Hypertext*. Kluwer Academic, Norwell, pages 27-47.

[2] Agosti, M., and Ferro N. (2003). Annotations: enriching a digital library. In Koch, T., and Sølvberg, I. T., editors, *Proceedings of the 7th European Conference on Research andAdvanced Technology for Digital Libraries (ECDL 2003),* LNCS 2769. Springer, Heidelberg, pages 88-100.

[3] 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. LNCS 3232, Springer, Heidelberg, Germany.

[4] Agosti, M. and Ferro, N. (2005). Annotations as context for searching documents. In Crestani, F. and Ruthven, I., editors, *Proceedings of the 5th International Conference on Conceptions of Library and Information Science - Context: Nature, Impact and Role*. LNCS 3507. Springer, Heidelberg, pages 155-170.

[5] Agosti, M., Ferro. N., and Orio, N. (2005). Annotating illuminated manuscripts: an effective tool for research and education. In: Marlino, M., Sumner, T. and Shipman III, F.M., editors, *Proc. 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005)*, ACM Press, New York, USA, 2005, pages 121-130.

[6] Agosti, M. and Ferro, N. (2006). Search Strategies for finding annotations and annotated documents: the FAST Service. In Larsen, H. L., Pasi, G., Ortiz-Arroyo, D., Andreasen, T., and Christiansen, H., editors, *Proceedings of the 7th International Conference on Flexible Query Answering Systems (FQAS 2006),* LNAI 4027. Springer, Heidelberg, pp. 270-281.

[7] 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.

[8] Agosti, M., Ferro, N., and Orio, N. (2007). Annotations as a Tool for Disclosing Hidden Relationships Between Illuminated Manuscripts. In Basili, R. and Pazienza, M. T.,

editors, *Proceedings of AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, 10th Congress of the Italian Association for Artificial Intelligence*, LNCS 4733, Springer, pages 662-673.

[9] 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.

[10] Barbera, N., Meschini, F., Morbidoni, C., and Tomasi, F. (2012). Annotating digital libraries and electronic editions in a collaborative and semantic perspective. In Agosti, M., Esposito, F., Ferilli, S., and Ferro, N., editors, *Digital Libraries and Archives. 8th Italian Research Conference (IRCDL 2012)*,. CCIS 354, Springer, Heidelberg, Germany, pages 46-57.

[11] Bélanger, M.-E. (2010, 02 03). Ideals. Retrieved 10 25, 2012, from https://www.ideals.illinois.edu/bitstream/handle/2142/15035/belanger.pdf?sequence=2

[12] Bhagwat, D., Chiticariu, L., Tan, W.-C., and Vijayvargiya, G. (2004) An annotation management system for relational databases. In: Nascimento, M. A., Özsu, M. T., Kossmann, D., Miller, R. J., Blakeley, J. A., and Schiefer, K. B., editors, *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004)*. Morgan Kaufmann, pages 900-911.

[13] 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/

[14] Buneman, P., Kostylev, E. V., and Vansummeren, S. (2013). Annotations are relative. In Tan, W.-C., Guerrini, G., Catania, B., and Gounaris, A., editors, *Proceedings of the 16th International Conference on Database Theory (ICDT 2013)*. ACM, New York, NY, USA, pages 177-188.

[15] 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.

[16] Deep Liquid, jCrop, http://deepliquid.com/content/Jcrop.html

[17] 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. LNCS 5714, Springer, Heidelberg, Germany.

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

[19] Fielding, R. T. and Taylor, R. N. (2002). Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology (TOIT)*, 2(2):115–150.

[20] Hallam-Baker, P. M. and Behlendorf, B. (1996). Extended Log File Format - W3C Working DraftWD-logfile-960323. http://www.w3.org/TR/WD-logfile.html.

[21] 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. LNCS 7616, Springer, Heidelberg, Germany.

[22] 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. CCIS 343, Springer, Heidelberg, Germany.

[23] Kahan, J. and Koivunen, M.-R. (2001) Annotea: an open RDF infrastructure for shared Web annotations. In Shen, V. Y., Saito, N., Lyu, M. R., and Zurko, M. E., editors, *Proceedings of the 10th International Conference on World Wide Web (WWW 2001)*, ACM Press, New York, pages 623-632.

[24] Kostylev, E. V. and Buneman, P. (2012). Combining dependent annotations for relational algebra. In Deutsch, A., editor, *Proceedings of the 15th International Conference on Database Theory (ICDT 2012)*. ACM, New York, NY, USA, pages 196-207.

[25] Lawless, S., Hampson, C., Mitankin, P., and Gerdjikov, S. (2013). Normalisation in Historical Text Collections. In *Proceedings of Digital Humanities 2013*, Lincoln, Nebraska, USA [In Press].

[26] Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In Akscyn, R., editor, *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HT 1998): links, objects, time and space-structure in hypermedia systems*, pages 40-49. ACM Press, New York.

[27] Martínez, J. M. (editor) (2004). *MPEG-7 Overview*. ISO/IEC JTC1/SC29/WG11N6828. Palma de Mallorca, Spain, pages 74.

[28] MPEG home page, http://www.chiariglione.org/mpeg

[29] Nagao, K. (2003). *Digital content annotation and transcoding*. Artech House Publishers.

[30] 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.

[31] Okfn. (n.d.). Okfn Annotator. Retrieved 06 2012, from http://okfnlabs.org/annotator/

[32] Rangy, http://code.google.com/p/rangy/

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

[34] Sanderson, R. and Van de Sompel, H. (2010). Making web annotations persistent over time. In Hunter, J., Lagoze, C., Giles, C. L., and Li, Y.-F., editors, *Proceedings of the 2010 Joint International Conference on Digital Libraries (JCDL 2010)*. ACM, pages 1-10.

[35] TILE. (2011). TILE: text-image linking environment. Retrieved 07 2012, from http://mith.umd.edu/tile/

[36] Trinity College Dublin, 1641 Depositions. http://1641.tcd.ie/

[37] Università degli Studi di Padova, IPSA (*Imaginum Patavinae Scientiae Archivum*). http:/ipsa.dei.unipd.it/en_GB/

[38] Wang, X., Ye, L., Keogh, E., and Shelton, C. (2008). Annotating historical archives of images. In Larsen, R. L., Paepcke, A., Borbinha, J. L., and Naaman, M., editors, *Proceedins of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2008)*, ACM, pages 341-350.