Facilitating Fine-grained Open Annotations of Scholarly Sources

Peter Boot
peter.boot@huygens.knaw.nl
Huygens ING, The Netherlands

Ronald Haentjens Dekker
ronald.dekker@huygens.knaw.nl
Huygens ING, The Netherlands

Marijn Koolen
marijn.koolen@huygens.knaw.nl
Huygens ING, The Netherlands

Liliana Melgar
melgar@uva.nl
University of Amsterdam, The Netherlands

Introduction

In the scholarly domain, annotation is a fundamental activity (Unsworth, 2000). Current web-based annotation facilities enable a specific way of annotation (via note-taking, highlighting or commenting) which are useful when scholars are exploring or gathering an initial set of resources, but more sophisticated support is needed for detailed analysis, close reading, and data enrichment. At this point, it is important to take into account the structural relations between documents and their parts. For example, when annotating a letter, annotation tools should be aware that a targeted text fragment is the name of the sender, or that the annotation of a film targets the intellectual work instead of the specific version or copy on which the annotation is made.

In addition, many standalone tools use annotation models with idiosyncratic solutions to enable the relations between different media objects and their parts, which limits the possibilities to exchange those annotations . In general, there is a lack of necessary details for durable access to and interpretation of annotations. For this, detailed information is needed about the annotated object, the annotator and the annotation itself (Melgar et al. 2016, Walkowski & Barker, 2010). In this paper we focus on the requirements for the annotated object, in a web-based environment, and propose a method for making necessary details of objects openly available for any annotation tool.

Requirements of scholarly annotation

In line with (W3C 2017b) we refer to the object that is annotated as the annotation target, the content of the annotation as the annotation body and who or what creates the annotation as the annotation creator. All three are complex entities with aspects that have consequences for interpreting an annotation (Melgar et al., 2016).

Annotation Creator: With respect to the creator it is important to know the intention/motivation for making the annotation (Walkowski & Barker, 2010) and when sharing and reusing annotations, their level of expertise, both in terms of the scholarly domain and in the nature of the annotation task (e.g. the amount of experience/expertise of the annotator in classifying objects according to a controlled vocabulary).

Annotation target: of the target it is important to know which part of the object is targeted. This is not merely about addressing media fragments. Media (e.g., html, mp3, jpg) are carriers of abstract information objects (scenes in movies, chapters in books, objects in pictures) with different conceptual levels (e.g. work, expression or manifestation , see Figure 1) and it is essential to be able to address those abstract objects and the relationships between them.

Annotation body: Of the content of the annotation it is important to know its nature (a natural language comment, structural or subject metadata, a link to another resource), in what form it is made (e.g. closed representation or natural language representation), at what level of control (from mostly uncontrolled to strictly controlled and structured) and for what scholarly purpose, e.g. gathering or exploring sources or thematic or stylistic analysis (Melgar et al., 2017).

State of the Art

There are various models for capturing digital annotations to make them accessible and interpretable. The Web Annotation Data Model (W3C 2017a, 2017b) is a generic model that covers aspects of the annotation body, target and creator. This model focuses on annotations in the context of online social

interaction (e.g., commenting, sharing), not necessarily on scholarly annotations done during analysis or data enrichment .

An extended model specifically for scholarly research was proposed by Hunter et al. (2011), which includes context aspects for both the annotation body and target. The Annotating All Knowledge Coalition is also directed at scholarly annotation and lists several issues, including:

1. The lack of support for discovery, sharing and reuse of annotations.
2. Underutilization of collections.
3. The closed and non-standardized nature of current annotation tools.

Current annotation support is either part of a suite of functionalities in monolithic applications with their own models for annotation (e.g. TextGrid , Textual Communities , eLaborate , CATMA for text,Elan and Anvil for multimedia materials, and QDA software packages for mixed media qualitative data analysis), or they lack specificity in describing the annotation target, e.g. Hypothes.is (Perkel, 2015) and Pundit (Grassi et al., 2012) and site-specific annotation tools, e.g. in The Diary of Samuel Pepys).

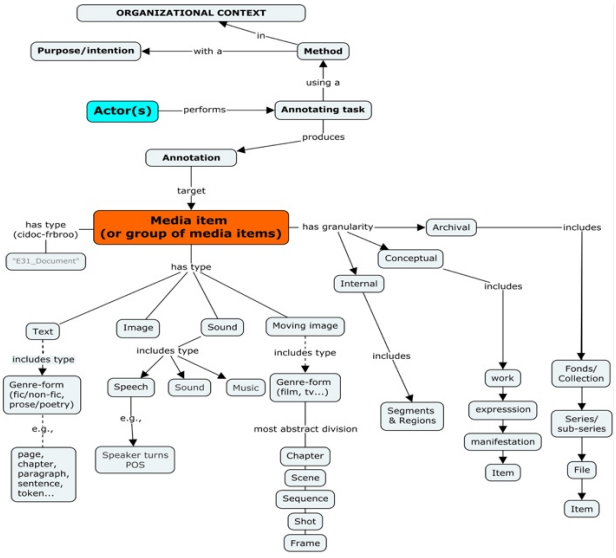

Figure 1. Conceptual model of annotated object (details of other parts of the model are left out for clarity)

Building on earlier work (Melgar et al., 2016), in this paper we argue the need for application support for more specificity of the annotation target (see Figure 1). We identify two additional issues with the current state-of-the-art:

4. The W3C annotation protocol lacks support for a potential annotation target identifying and describing itself to the annotation tool.
5. The model also lacks a schema, which would allow scholars or website maintainers to define constraints for a specific class of annotations that is applicable in the context of a specific group of scholarly objects.

Use case: annotation in scholarly digital editions

These issues are illustrated by a scenario of a digital scholarly edition where scholars have a need for annotation support (Boot, 2009, Robinson, 2004, Siemens et al., 2012). Consider an edition that wants to incorporate an external annotation tool into its pages (Figure 2): an edition server shows an edition to a client in a browser. The annotation client runs within that same browser window, but doesn't know about the edition's structure and it talks with its own server. To communicate intelligently with the user, the annotation client needs information about the structure of the edition, which has to be provided by the edition.

The annotation tool should know about the edition's structure for a number of reasons:

- The edition often contains multiple representations of the same text fragment. There might be a diplomatic and a critical transcription, one or more translations, audio versions, and who knows what other versions, and annotations made in one of these should be available in others;
- Other sites may have other editions of this particular text. It should be possible to exchange annotations between them;
- The edition has an internal structure, e.g. a book divided in chapters, or the fragments appearing in modern authors' drafts, or the elaborate structure with multiple apparatuses of some editions of medieval texts. An annotation that refers to a specific component of an edition should be able to address that component and know what sort of component it is.
- The edition should be able to propose suitable annotation types for its components. For personal names, it might suggest an annotation type that links the person to an authority file. For transcriptions, there might be special

annotation types for proposed corrections to the transcription. Edition collaboratories could use the annotation functionality to solicit multiple sorts of specialised information from its collaborators.

This proposal requires that: (i) the edition describes itself and its structure to the annotation tool, and provides suitable labels for the annotatable objects; (ii) the edition can suggest annotation types for the annotatable objects; (iii) the effort to integrate annotation functionality in existing editions is minimal; (iv) the annotation tool is generic, but able to handle the created annotations with awareness of the structure that they apply to (it can e.g. return aggregated annotations); (v) the annotation targets are durable and not formulated in terms of HTML structure; and (vi) URI's should be treated as opaque (i.e., we shouldn't try to guess the relations between the annotated components based on their URIs); and lastly (vii) URIs should be canonical.

Proposed Solution

We propose a solution similar to Schema.org (an initiative for adding structural semantics to information on the web) whereby descriptive information about annotatable resources is made accessible to the client by embedding it in the HTML presentation layer through RDFa attributes (Figure 3), using an extensible resource descriptive ontology. Figure 4 shows a basic ontology for text objects (left half of Figure 4) with an edition-specific extension for the example edition (right half of Figure 4). This ontology shares concepts with the FRBRoo ontology (Bekiari et al., 2015) but starts from specific annotation-related concepts. In future work we will investigate extending the ontology with FRBRoo concepts.

Although this approach is focused on annotation of resources on the web, the same principle could be applied in offline annotation, if the offline resources are described in a similar way and annotation clients are developed to make use of this. Also, descriptive information for textual sources can be embedded as markup, but for audiovisual documents, this has to be done via a separate representation, for instance using SMIL (Bulterman et al., 2008).

```
<body vocab="http://huygens.knaw.nl/ns/annotate#" about="um:vangogh:let633"
typeof="CreativeWork">
  <span property="hasType" content="Letter"/>
  <h4 resource="um:vangogh:correspondence" typeof="CreativeWork" property="isPartOf">
    <span property="hasType" content="Correspondence">Van Gogh. The Letters</span>
  </h4>
  <p resource="um:vangogh:let633:par.1" typeof="CreativeWork" property="hasPart">
    <span property="hasType" content="Paragraph"/>Mon cher Bernard –
  </p>
  <p resource="um:vangogh:let633:note.1" typeof="Enrichment" property="hasEnrichment">
    <span property="hasType" content="Note"/>
  <p resource="um:vangogh:let633:page.1" typeof="TextBearer" property="isCarriedOn">
    <span property="hasType" content="Page"/>
```

Figure 3. HTML fragments of a letter of Vincent van Gogh (http://vangoghletters.org/orig/let633) described by embedded RDFa. The letter is identified by a URN (urn:vangogh:let633) and defined as a CreativeWork. It is part of a larger CreativeWork, Van Go

Figure 4. Basic ontology for text objects (left of dashed line) and extended ontology for Van Gogh Letters Collection (right of dashed line). The basic ontology recognizes three types of annotatable things: the creative work being edited and its parts (also creative works), the text bearers (e.g. manuscript pages), and editorial enrichment of any sort. Projects can create an extended ontology to suit their needs. The extended ontology shown here creates specialized classes for the needs of the Van Gogh letter edition (http://vangoghletters.org/).

Methodological impact

In our proposal annotatable resources describe their own semantic structure, thereby facilitating fine-grained annotations. With the RDFa attributes, annotation clients can identify the annotation target in terms of the resource structure (issue 4), which makes annotations less dependent on specific views on the underlying object. Furthermore, this allows development of lightweight open source annotation clients that web services can easily embed to bring annotation to collections of scholarly interest (issue 3).

This makes it easier for scholars to use and reuse annotations to support the argument made in a scholarly article (issue 1). It allows distinguishing different groups of annotations, so researchers can choose to display certain groups of annotations, thereby avoiding being drowned by irrelevant annotations (issue 5). It facilitates employing annotation functionality to ask for targeted comments on resource parts (what do you think of this

translation? What clarification of this material are you missing?). Scholars can also more meaningfully combine and compare them across collections and media types, e.g. analyse the correspondence between book and film versions of an intellectual work (issue 2).

If the annotations are consistently stored using open protocols, it becomes possible to reference them in scholarly publications. Collateral benefit of floating this form of 'deep web' semantics to the surface is that other external services such as search engines can also use the exposed semantic information to reason about available resources.

Bibliography

Bekiari, C., Doerr, M., Riva, P., Le Bœuf P. (2015). FRBR, object-oriented definition and mapping from FRBRer, FRAD and FRSAD - International Working Group on FRBR and CIDOC CRM Harmonisation, Version 2.4, November 2015.

Boot, P., (2009). Mesotext: digitised emblems, modelled annotations and humanities scholarship. Amsterdam University Press, 2009.

Bulterman, D., Hansen, J., Cesar, P. et al. (2008). Synchronized Multimedia Integration Language. W3C Recommendation 01 December 2008. https://www.w3.org/TR/2008/REC-SMIL3-20081201/

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., Ledda, G. (2012). "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries." SDA. 2012.

Hunter, J., Cole, T., Sanderson, R., van de Sompel, H. (2010). The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations. Presented at the Digital Humanities 2010. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-860.pdf

Melgar, L., Blom, J., Baaren, E., Koolen, M., Ordelman, R. (2016). A conceptual model for the annotation of audiovisual heritage in a media studies context. Presented at the AudioVisual Material in Digital Humanities 2016 workshop, Krakow, Poland. Retrieved from https://avindhsig.wordpress.com/workshop-2016-krakow/accepted-abstracts/liliana-melgar-jaap-blom-eva-baaren-marijn-koolen-roeland-ordelman/

Melgar, L., Koolen, M., Huurdeman, H.C., Blom, J. (2017). A Process model of Scholarly Media Annotation. In Proceedings of the 2017 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017.

Perkel, J. M. (2015) "Annotating the scholarly web." Nature 528.7580 (2015): 153-154.

Rizkallah, É., (2016). QDA software compatibility: Towards an exchange format with developers for their users. Presented at the Reflecting on the future of QDA software, Rotterdam, The Netherlands.

Robinson, P., (2004). "Where we are with electronic scholarly editions, and where we want to be." Jahrbuch für Computerphilologie Online 4 (2004): 123-143.

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., Sloetjes, H. (2008). An Exchange Format for Multimodal Annotations. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech.

Siemens, Ray, Timney, M., Leitch, C, Koolen, C., Garnett, A. (2012). "Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media." Literary and Linguistic Computing 27.4 (2012): 445-461.

Sloetjes, H. (2014). ELAN: Multimedia Annotation Application. In The Oxford Handbook of Corpus Phonology. Retrieved from http://www.oxfordhandbooks.com.proxy.uba.uva.nl:2048/view/10.1093/oxfordhb/9780199571932.001.0001/oxfordhb-9780199571932-e-019

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this. In Humanities Computing: formal methods, experimental practice symposium, King's College, London.

W3C, (2017a). Web Annotation Data Model. W3C Recommendation 23 February 2017. https://www.w3.org/TR/annotation-model/

W3C, (2017b). Web Annotation Vocabulary. W3C Recommendation 23 February 2017. https://www.w3.org/TR/annotation-vocab/

Walkowski, N-O., Barker, E.T.E., (2014). Digital Humanists are Motivated Annotators. Presented at the Digital Humanities 2014, Lausanne, Switzerland. Retrieved from http://dharchive.org/paper/DH2014/Paper-296.xml