Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context

John Bradley, Michele Pasin (DDH, King's College London)
{john.bradley, michele.pasin}@kcl.ac.uk

The emergence of formal ontologies into the World Wide Web has had a profound effect on research in certain fields. In the Life Sciences, for example, key research information has been captured in formal domain ontologies, like those mentioned in the Open Biological and Biomedical Ontologies website (OBOFoundary 2012). In parallel with this has been the development of the AO annotation ontology framework (AO 2012) which formalises annotation to connect ontologies such as those in the OBOFoundary to references to them in the scientific literature: an act sometimes referred to as "semantic annotation", and tools such as the SWAN annotation system (SWAN 2008) have emerged to support this. We will call the activity of linking references in a domain literature directly to entities in one or more domain ontologies "direct semantic annotation". We show it in schematic form in figure I. The annotations – shown as heavier lines connecting spots in the literature to the ontologies would be in the AO annotation ontology, or something similar to it.

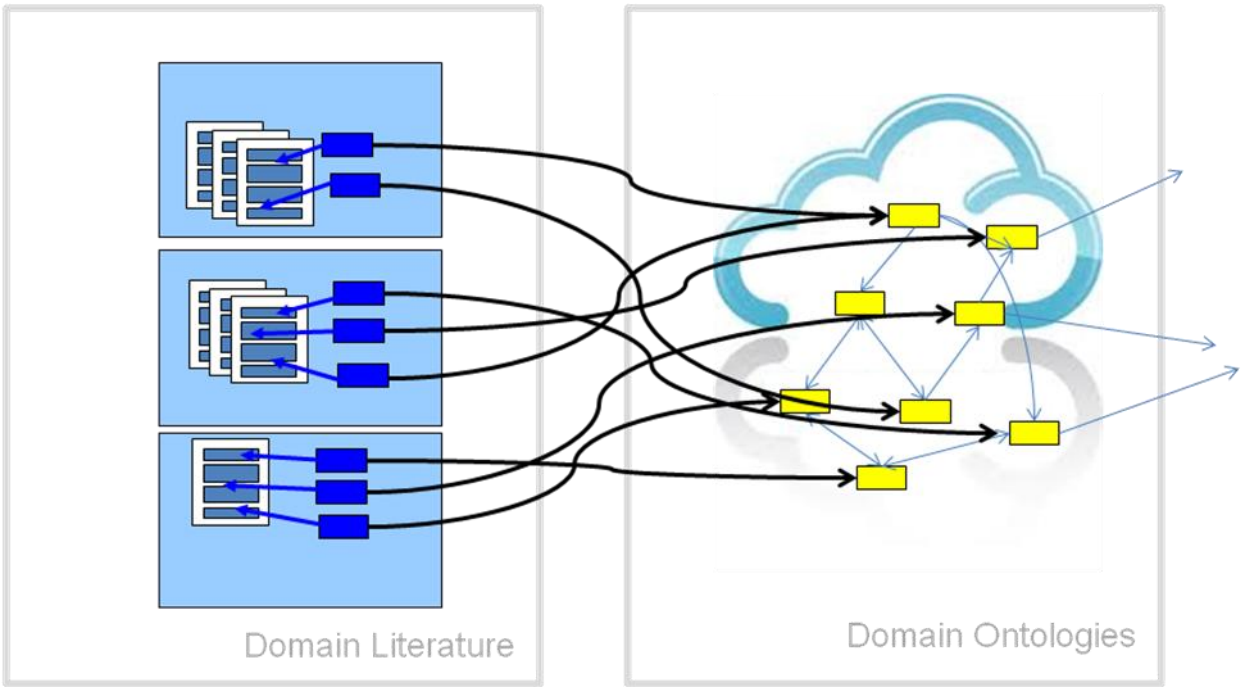

Figure 1: direct semantic annotation

Can direct semantic annotation be applied to research in the Humanities? For it to work as it does in the Life Sciences, formal models of humanities materials, such as CIDOC-CRM, need to exist and be already used to model material of interest to the humanities. Not much of this has happened at present, although perhaps Linked Data initiatives (Heath 2011) show some promise in that general direction.

Our department (DDH) has experience with projects that develop formal models that are something like ontologies. We have, for example, developed a formal structure for

prosopography that has worked well in projects such as the Paradox of Medieval Scotland prosopography (POMS 2011), and you can see our attempts to connect this to formal ontologies in Bradley and Pasin 2011. Other DDH projects also recognise the need for formal entities and have applied the Entity management EATS software (EATS 2011). Our Schenker project (Schenker 2012), which publishes the personal writing of the prominent musicologist Heinrich Schenker. is a case in point. EATS is used to manage historical entities such as persons, places, musical topics and sources, and contains a tool that allows one to perform direct semantic annotation to link references in Schenker's texts to these entities in its entity store. One highlights a textual reference with oXygen and uses the EATS tool to find the entity you want, or create a new one. The plugin then links the text to it by introducing a TEI *rs* tag (see figure 2).

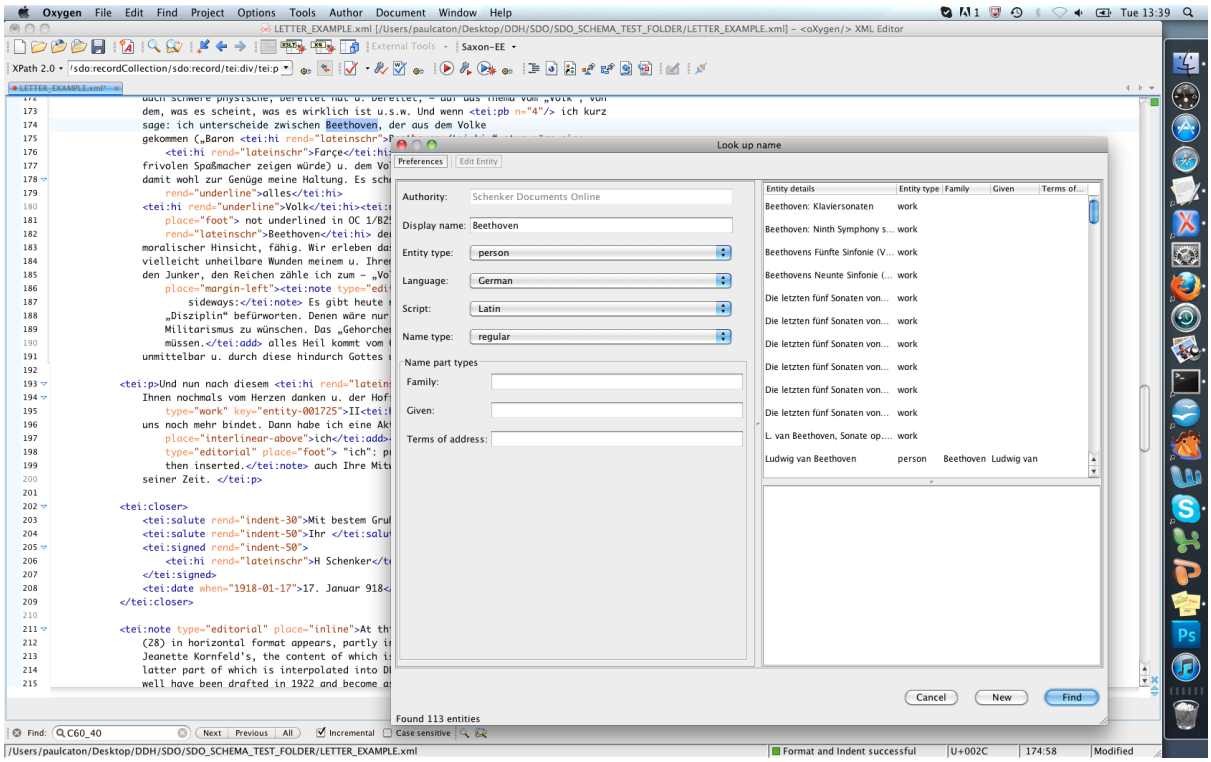

Figure 2: Semantic annotation with EATS in the Schenker Project

Mainstream Humanities Research: something different

Although direct semantic annotation to a pre-existing formal data model may be a key activity in the Schenker project, it does not represent the only kind of scholarly activity with texts. Indeed, most scholarship in the Humanities is not centered on this kind of labelling activity at all. Instead, almost all humanities scholars spend their time developing their *own* original interpretation of the materials they study, and aim to explore new concepts and paradigms about them which they present in their articles and books. (see Brockman *et* al 2001 and in Palmer *et al* 2009) The scholarship does not start out with predefined formal structures, but begins with a set of vague notions and insights in the scholar's mind as they read that only over time emerge clearly enough to be described in published work. Indeed, many in the humanities believe that clear, classical, thinking of the kind that is compatible with computer ontologies is incompatible with the kinds of things they want to say. See McGann 2004 for some thinking based on current humanities critical theory. We believe that his reference here to the radical

nature of quantum physics in the often-quoted phrase "hem of the quantum garment" (p 201) is not accidental.

Perhaps we need not go as far as McGann in this kind of deconstruction of humanities scholarship. However, we must recognise that for most humanists (a) scholarship is normally personal, (b) that it is meant to produce original ideas that must first emerge and then mature over time, and (c) that even when the ideas are mature enough for publication, they represent a model that is at least "pre-ontological", and perhaps at best only partly compatible with ontological modelling. When s/he writes an article a scholar does indeed wish to "create a model in the mind of the reader" (quoting the *NeDiMaH* workshop description) of their scholarly work, but their "model" may be only partially compatible with formal ontologies.

Pliny and formal models of personal scholarship

The Pliny project was launched by one of us to explore how computing could facilitate traditional scholarly practice. Pliny tried to be "Englebartian" – referring to Douglas Englebart's H-LAM/T paradigm (Englebart 1962) that successful software integrates with the human way of doing intellectual things so well as to almost disappear, and that this disappearing software can, paradoxically, sometimes allow its users to do entirely new things that they had been previously incapable of doing. Many researchers, including Brockman and Palmer mentioned earlier, have noted the importance of notetaking and management in humanities research. Thus, Pliny started with this at its centre, and modelled its approach on strategies for taking and managing notes as they were described in books like Altick and Fenstermaker's *The Art of Literary Research* (1992).

Out of this work came two models: the interface which developed a particular view of how users might usefully interact with a notetaking and management tool to help them develop their own interpretation, and the data model that stored the information. Bradley 2008 describes Pliny's user interface in terms of affordances: 2-dimensional space, containment and hierarchy, naming and labelling and multiple reference of notes material in different contexts, including typing of reference. In figure 3 we see Pliny's approach to its data: not only representing annotation, but also supporting and to some extent representing the development of original concepts.

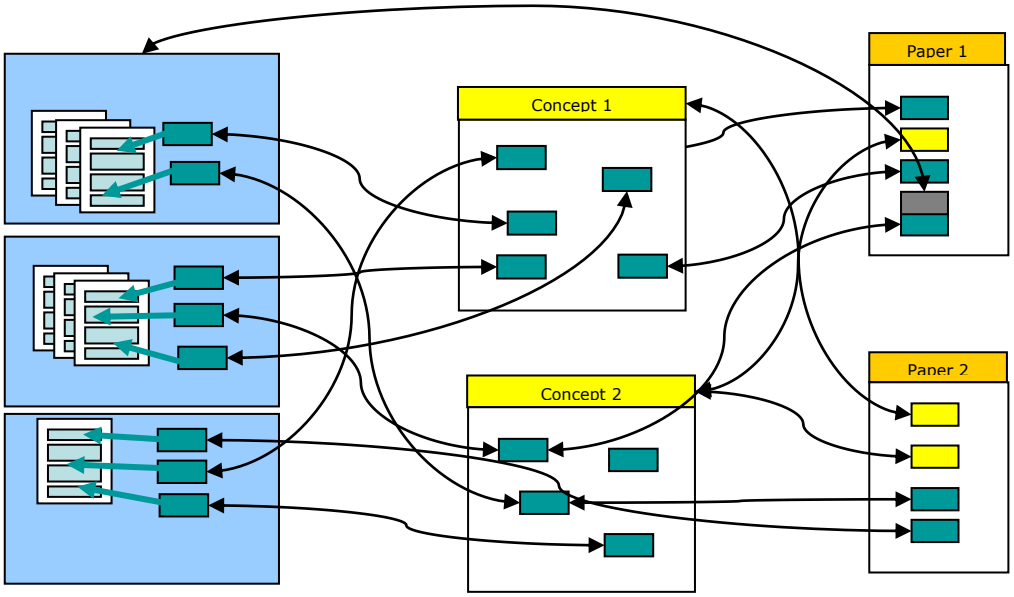

Figure 3: Pliny objects in its "notetaking" application (from Bradley 2012)

Here, the annotation of materials the researcher reads begins the research process (shown at the left), and traditionally ends with the writing of papers (right end). In between Pliny helps its user turn his/her observations into new, personal, conceptual structures (only 2 "concepts" are shown here, but a real user would have many more), which can then be used to develop ideas for papers. Note the difference from direct semantic tagging in figure I: the annotations do not link directly to pre-existing formal ontological entities, but first appear as informal prose notes that may, as the researcher's understanding grows, acquire a more formal representation and emerge as entity-like objects in the form of personally developed new concepts, themes, ideas, etc.

Pliny's data model is strongly suggestive of RDF and other base ontological technologies. Like RDF, the structure is a network and the links between the network nodes can be typed in a way similar to a RDF predicate. Pliny comes with the ability to export its structure into a Topic Web format, and some preliminary work has been done (see Jackson 2010) to map Pliny data into RDF through the Open Annotations Collaboration (OAC 2011) ontology. Adding mechanisms to Pliny's interface to allow a user to link items via RDF from their personal model to external formal entities such as, say, CIDOC-CRM is certainly possible.

The resulting paradigm is one that, unlike direct semantic annotation, separates the annotation of the domain literature from the highly formal world of domain ontologies by injecting a personal interpretative component in-between. One introduces a personal, more informal, representation of the scholarship into the picture, and this in ways that are compatible with RDF.

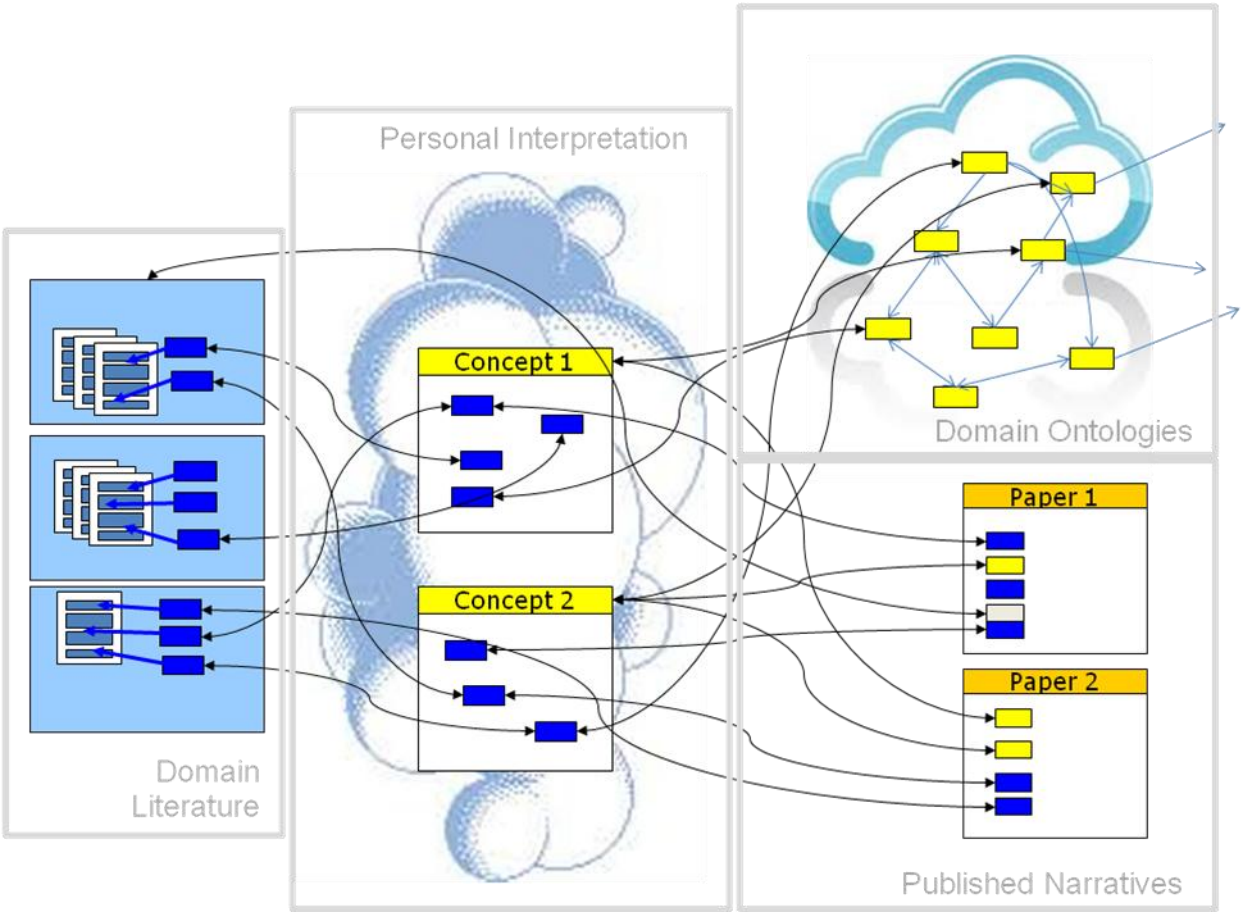

Figure 4: The place of a structured personal Interpretation

Figure 4 is similar to the direct semantic annotation model shown in figure 1, but adds a structured object representing the personal interpretative work of an individual that sits between a scholar's reading material, and any shared public domain ontologies. This personal interpretation "cloud" might well never be a clear-cut as formal ontologies must be, but it's presence here recognises and enables the *process* towards formality that is a central part of interpretation in humanities scholarship. By interposing this somewhat-informal semantic "cloud" between the texts and the formal ontologies of the semantic web, we see a way of thinking about this central personal interpretive work that fits with the larger, more formal, semantic web world. Although the nature of traditional humanities research does not suit the standard direct semantic annotation model of the Life Sciences, Pliny provides an approach that, over time, encourages the researcher to turn this cloud of personal interpretation into material that becomes more and more compatible with computer ontologies and the semantic web.

References

Altick, Richard D and John J Fenstermaker (1992). *The Art of Literary Research*. New York: W.W. Norton & Company.

AO (2012). *AO: Annotation-ontology.* Website at http://code.google.com/p/annotation-ontology/

Bradley, John (2008). "Thinking about Interpretation: Pliny and Scholarship in the Humanities", *Literary and Linguistic Computing* 2008 Vol. 23, No. 3 pp. 263-79. doi: 10.1093/llc/fqn021. Online at http://llc.oxfordjournals.org/cgi/reprint/fqn021?ijkey=3UzJDubDB0FRQcR&keytype=ref

Bradley, John (2012). "Beyond Digital Media: Moving beyond a 'media' orientation in the annotation of digital objects". Accepted for publication in the *Digital Humanities Quarterly*. A draft (under a different name) of this article is available at http://pliny.cch.kcl.ac.uk/docs/article-2011.pdf

Bradley, John and Michele Pasin (2011). Prosopography and Computer Ontologies: towards a formal representation of the 'factoid' model by means of CIDOC-CRM. Given as a part of DDH's KR research seminar. Online at http://www.slideshare.net/mpasin/prosopography-and-computer-ontologies-towards-a-formal-representation-of-the-factoid-model-by-means-of-cidoccrm.

Brockman, William S., Laura Neumann, Carole L. Palmer, Tonyia J. Tidline (2001). *Scholarly Work in the Humanities and the Evolving Information Environment, a report from the Council on Library and Information Resources* (Washington DC: Digital Library Federation, Council on Library and Information Resources). (2001_. Online version at http://www.diglib.org/pubs/dlf095/

EATS 2011. *eats: Entity Authority Tool Set: a web application for authority control*. Website at http://code.google.com/p/eats/

Englebart, Douglas. (1962). Augmenting Human Intellect: A conceptual framework. Stanford CA: Stanford Research Institute. Online at http://www.bootstrap.org/augdocs/friedewald030402/augmentinghumanintellect/AHI62.pdf (accessed March 2007)

Jackson 2010. *RDF-encoding Pliny annotations in the Open Annotation Collaboration project*. University of Illinois: GSLIS Technical Report #ISRN UIUCLIS--2010/2+OAC.

McGann, Jerome (2004). “Marking Texts of Many Dimensions”. In Susan Schreibman, et al. *A Companion to Digital Humanities*. Oxford: Blackwell. pp. 198-217.

OAC (2011). *Open Annotation Collaboration*. Website at http://www.openannotation.org/

OBOFoundary 2012. *The Open Biological and Biomedical Ontologies*. Website at http://obofoundry.org/.

Palmer, Carole L., Lauren C. Teffeau, and Carrie M. Pirmann. 2009. *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Report produced by OCLC Research. Available online at: http://www.oclc.org/research/publications/library/2009/2009-02.pdf (.pdf: 412K/59 pp.).

POMS 2011. *Paradox of Medieval Scotland: 1093-1286.* Website at http://www.poms.ac.uk/

Schenker 2012. *Schenker Documents Online*. Website at http://www.schenkerdocumentsonline.org.

SWAN 2008. *SWAN Project: Semantic Web Applications in Neuromedicine*. Website at http://swan.mindinformatics.org/.