The Landscape of Digital Annotation and Its Meaning

Niels-Oliver Walkowski*

* TELOTA, Berlin-Brandenburg Academy of Sciences and Humanities
Jägerstraße 22/23, 10117 Berlin
walkowski@bbaw.de

Abstract

In this paper I argue that due to the development of computational environments the usage scenarios and the interpretation of annotations have become an increasingly complex issue. Furthermore, I evaluate the different contextual dimensions which constitute the meaning of annotation data and present a formal scheme for their systematization.

1. The Rise of Annotations

The topic of annotations has raised significant interest in the field of Digital Humanities and beyond during the last years. Many ESFRI[1] projects from the Humanities or Social Sciences have worked on annotations in corresponding working packages. With the pund.it[2] EUROPEANA[3] has also built its own annotation tool. In the W3C two working groups developed a web compliant standard model for annotations (Sanderson, Ciccarese, & Van de Sompel, 2013) and specifications for a web annotation architecture[4]. Finally, projects like annotator.js[5] and hypothes.is[6] gained tremendous community interest as well as significant funding. All these activities have significantly multiplied the usage and usage scenarios of annotations.

How do these activities affect the understanding of annotations or the ways of annotating? There are two possibilities to read the overambitious title of this study. The first version asks for the meaning of the landscape itself that is to say what is the meaning of the way this landscape is shaped. For instance, what does the structure of this landscape tell us about the state of annotations in the digital age. The second version refers to the meaning that is represented in annotations itself and which forms part of this landscape. To put it more concisely, the question here is: what do all these annotations mean and how do we find out?

The work which is presented in the current paper was financed by DARIAH-DE and benefits from a survey that was carried out by the 'DARIAH-EU Working Group Digital Annotations' (DARIAH-EU, 2016). Hence, the question is also how does this development challenge the work of infrastructure projects and why should infrastructure projects take care about these questions.

Infrastructure projects are building storage, services and tools among other things. As I mentioned before, this holds especially true for annotation data at the moment.. However, as both Atkins (Atkins, 2003) and Rockwell (Rockwell, 2010) points out infrastructure needs to be more to become successful. It needs to also take care about the research ecology which it seeks to serve. In the case of annotations this means to be informed about new digital annotation practices and to assure that annotations are used in a sound way. The dynamic which has been described before as well as the peculiar nature of annotation data makes this task especially challenging for annotations. They are mostly a granular, highly context dependent piece of information.

Another reason is given by the fact that DARIAH is a project in the (Digital) Humanities that means embedded into the Humanities research tradition. In the Humanities research about the modes of knowledge production itself is a crucial part of the research portfolio. Additionally, annotating has a long tradition in humanist research practice. Thus, research on digital annotations is also a great chance to push forward the integration of digital methods in the Humanities as Meister (2015) points out.

Having all this said, the title of this paper could be transformed into two questions:

- What needs to be known from annotation contexts so that annotation data can be reasonably used elsewhere? In technical terms, what are the metadata needs?
- What are annotations today or in the language of infrastructure projects, which best practices in annotating exist?

In the rest of this paper I will try to come closer to an answer for these questions.

2. The Meaning in Annotations

For the purpose of providing rich information to support the evaluation of these questions the DARIAH-EU Working Group Digital Annotations developed a sophisticated questionnaire with over 40 questions. The main goal of the survey is not to derive statistical claims but to generate a dense description that represents the complexity of the topic. Thus, the number of investigated use-cases is relatively low and include 17 filled out questionnaires.

Before getting into detail, a general result of the survey addresses the value of the whole effort as such. More precisely, it was asked if project members think that annotations from their work could be fully understood on its own and without further knowledge about the project. A positive answer to this question seems appropriate in a data sharing context

1 The European Strategy Forum on Research Infrastructures is a policy making body of the European Commission for the development of (digital) research infrastructures
2 http://thepund.it/
3 http://www.europeana.eu/portal/
4 https://www.w3.org/annotation/
5 http://annotatorjs.org/
6 https://hypothes.is/

Curiously enough, the response pattern resembles a common pattern of similar questionnaires regarding research data sharing in general. Only one person answered negatively while another one abstained. However, the number of purely positive answers were also only four. The majority of people wrote 'Yes, but …'. I do not only interpret this result as a clear expression in favour of further research on metadata needs for sharing annotation data but also for uncertainty about the status of annotations in terms of data sharing.

2.1. Meaningful Dimensions of Annotations Identified in the Survey

Next, I will give a short overview about important different dimensions for the meaning of annotations in the light of the survey. Some of these dimensions were explicitly mentioned by the participating projects. Others were extracted from answers where they implicitly address issues of meaning construction.

The first aspects which influences the appropriate interpretation of the meaning in annotations concerns the technological production of the annotated object. In a use-case from the field of visual anthropology[7] the participant remarks that knowledge about the process of ethnographic film making is very supportive to understand the annotations about these films. Likewise, the Monasterium[8] use-case reveals that certain annotations on documents can only be understood with knowledge about the creation of digital copies of these documents.

In the DARIAH-DE Fellowship use-case a comparable issue is mentioned but evaluated slightly different. This use-case addresses the issue of annotations in which their content might not be enough for its interpretation. However, the investigation of the annotated object region provides sufficient context information. The link between both might seem obvious. However, in digital annotating the target might not be at the same place as the annotation body. This can cause problems of different types. Dereferenceability and even more renderability of annotated objects should therefor be a crucial aspect.

Another dimension does also concern the annotated object. However, this time it is about the question what is technically referenced. The Video Annotation in Transcultural Studies[9] (VATS) as well as the Semantic Topological Notes[10] (SemToNotes) use-case emphasize that many annotation services do not offer the possibility to exactly reference a shape in an image. Instead, the annotation reference creates a box around the shape of interest. This is not precise and can lead to information retrieval and interpretation issues. We can call this the fragment dimension.

The fragment dimension is part of a bigger issue. This issue is about the concrete object layer which is addressed by the annotation. Some examples will clarify what is meant. The Relations in Space use-case in which inscriptions in Jewish gravestones are annotated distinguishes between annotations about the carrier of the inscription (the gravestone) and the inscription itself. Accordingly, annotations which reference the same area in the digital copy address completely different facets.

In the Monasterium illuminated areas in the digital copies are annotated next to layout information. Thus, the first group of annotations do not address the material nor some basic semantic concepts (title, paragraph, among others), they describe aspects of the digitization.

In e-Metaphor annotating a part of text as a metaphor does not just mean the metaphor itself but the 'focus and frame of metaphorical construction'. Within the illustrative terminology from literature studies it addresses an intratextual dimension of the metaphor and not just the metaphor.

Alluding to a common term in the research field of Systemic Functional Grammar the context dimension which discriminates the annotated objects in levels between materiality and intertextuality can be called the strata dimension.

Both the use-case Visual Anthropology as well as Ethnomusicology[11] highlight that knowledge about the way Ethnologists or Musicologists work and process content are supportive facets to understand well corresponding annotations. This dimension is the methodology or practice dimension.

Another meaningful dimension for the sound use of annotations might seem too obvious and trivial to consider. However, it is a very important dimension. The form and properties of the annotation itself needs to be clear. To put it in technical terms, the model needs to be transparent. I mentioned the standardized Open Annotation Data Model before. Nevertheless, the fact that such a model exists does not mean that it is always used or can be used everywhere - technically as well as semantically.

For instance, the e-Codicology[12] use-case has defined its annotation model in a proprietary SKOS model. The VATS use-case creates attention for fact that the annotation body can be encoded in a way that needs information about what is required to render it. This can be something like MIME-Type information for example. On a semantic level the DBPedia Spotlight use-case indicates that it needs an explanation of a specific property called the 'popularity score'. This dimension is the model dimension of annotations.

Knowledge about structure and properties of annotations is one thing, knowledge about concepts which are used in annotations are another. Certainly, this issue is well discussed in the Semantic Web domain and a lot of annotation data is produced following the Semantic Web compliant cause of conduct. Nonetheless, this issue is a lot more complicated and numerous examples in the survey give evidence about this fact.

For instance, e-Codicology uses TEI but in many cases this information is not sufficient and encoding principles are necessary. The e-Poetics use-case applies a technical terminology which complies with the rules of literature study. The issue is that no technological representation for this terminology exist. The example of e-Metaphor is even more complicated. E-Metaphor's concept of metaphors is defined very precisely for the purpose of the project. In this case it is the peculiarity of the definition linking to a

7 http://isn3.zrc-sazu.si/avl_arhiv/index.php (registration required)
8 http://www.monasterium.net
9 http://vad.uni-hd.de/
10 http://hkikoeln.github.io/SemToNotes/
11 http://etnofletno.si/
12 http://www.ecodicology.org/

specific type of theory of metaphors which complicates the appropriate use of annotations and which might create misunderstandings. Having all this said, the problem of the semantic dimension of annotations goes far beyond the question if a formal and technical representation of its concepts exist.

An interesting dimension of annotations has already been investigated and named very well by Agosti, Bonfiglio-Dosio, & Ferro (2007). In the cited work the authors remark that the meaning of annotations is often shaped by relations between annotations of the same annotating process. They call this the 'dialogic' aspect of annotations. The use-cases DHWork and Visual Anthropology also highlights this aspect in some of their answers. Accordingly, annotations should always contain information which make it possible to dereference corresponding annotations..

The correct angle to understand annotations is often set by knowing the purpose of an annotation process together with the research goals. There are no better examples for this link than the use of annotations in e-Metaphor and in the DARIAH-DE Fellowship use-case. In both cases annotations are produced in a training process of mining algorithms. As documentation for the development of this algorithm these annotations are incredibly interesting. However, as serious annotations about the annotated object part of them do not serve well.

In the CATMA[13] use-case annotations are created in a crowdsourcing environment. They are meant to be heterogeneous for the purpose to engender a dense description. Thus, their function must not be interpreted as normative classification. Finally, the DHWork use-case remarks that one of its goals is to evaluate the difference between annotation and comment. This distinction shapes what information might be published as annotation data and what is not. Thereby, it puts specific information in a specific context which depends on the definition of annotation. Thus, annotation should reference the results of the research process in which they were created.

The last dimension which significantly shapes the meaning of annotations is their intended audience. This phenomenon was intensively discussed by Chiang (2010). Accordingly, form and content of annotations differ when they are produced to support an individual researcher, a research group or meant to be public. The Visual Anthropology use-case highlights a scenario in which this issue is quite obvious. Annotations in correspondence with field diaries in ethnology often contain information which are ethically problematic. However, the issue exists also in more subtle scenarios.

2.2. Evaluation

Several approaches have already been carried out to systematize context dimensions of digital annotations into consistent models. Before I introduce my own approach I would like to quickly outline the drawbacks that these attempts still possess.

One problem of comparable approaches is that they are often carried out on the ground of settled annotation scenarios. For instance, Chiang (2010) develops a sophisticated 'Annotation Function Coding Scheme' based on 'A Multi-Dimensional Approach to the Study of Online Annotation'. However, in her research online annotations are annotations produced in the interpretative reading process of text based web documents. Thus, the work tackles an individual annotation scenario which is relatively well known.

In other cases the research which is carried out only focus on specific dimensions of annotations. Likewise Bauer & Zirker (2015) tackle the problem of different interpretation levels (called strata before) while Bélanger (2010) or Gradmann et al. (2015) concentrate on the relationship between research practice and annotations.

Approaches like oa:Motivation in the Open Annotation Data Model albeit improved over time by adding oa:hasPurpose to oa:motivatedBy are still limited, inconsistent and contingent as I have argued elsewhere (Walkowski, 2015, 2016). This might relate to the fact that originally motivation was included into the model to provide interesting ways of querying annotation data[14]. The issue of correct interpretation and usage of annotations was not the driving force.

Other approaches which consider a variety of computational annotation scenarios like Agosti et al. (2007) make transparent the complexity of the issue but do not intent to get into greater detail. Finally, many approaches only tackle the topic in prose but not in a formal manner.

In this paper I want to introduce a first attempt to systematize the different context dimensions of annotations which were addressed at least partially in the survey before. For this purpose, I would like to introduce the diagram presented in figure 1.

In the center of the figure there is the annotation which consists of a body and a target. The target addresses both the annotated part of the object as well as the entire object. The body holds the annotation content. The upper half of the figure addresses aspects of practice and semantics while the lower half references aspects of technology and structure.

Furthermore, a production flow exist from the left to the write in which objects for annotations are created, then annotated and annotations are processed. The left half represents both the production of an object to be annotated and the annotation itself. Likewise, goal and publication can identify corresponding activities which belong to annotations or output for which a peculiar annotation is created. Finally, two types of relationships exist between annotation target and annotation body. Depending on the situation, each dimension can be formally instantiated in a way that expresses its contribution to the overall meaning of peculiar annotation.

For instance, the revision of annotations in the visual anthropology use-case caused by ethical concerns is part of the publication dimension while algorithm testing in the e-Metaphor use-case provides a goal dimension. However, the dimension can also be more obvious. A semantic tag which is taken from a RDF taxonomy and references this taxonomy by namespace completely opens up the semantic dimension in annotation bodies.

13 http://www.catma.de/

14 refer to the project wiki for further details at https://www.w3.org/community/openannotation/wiki/

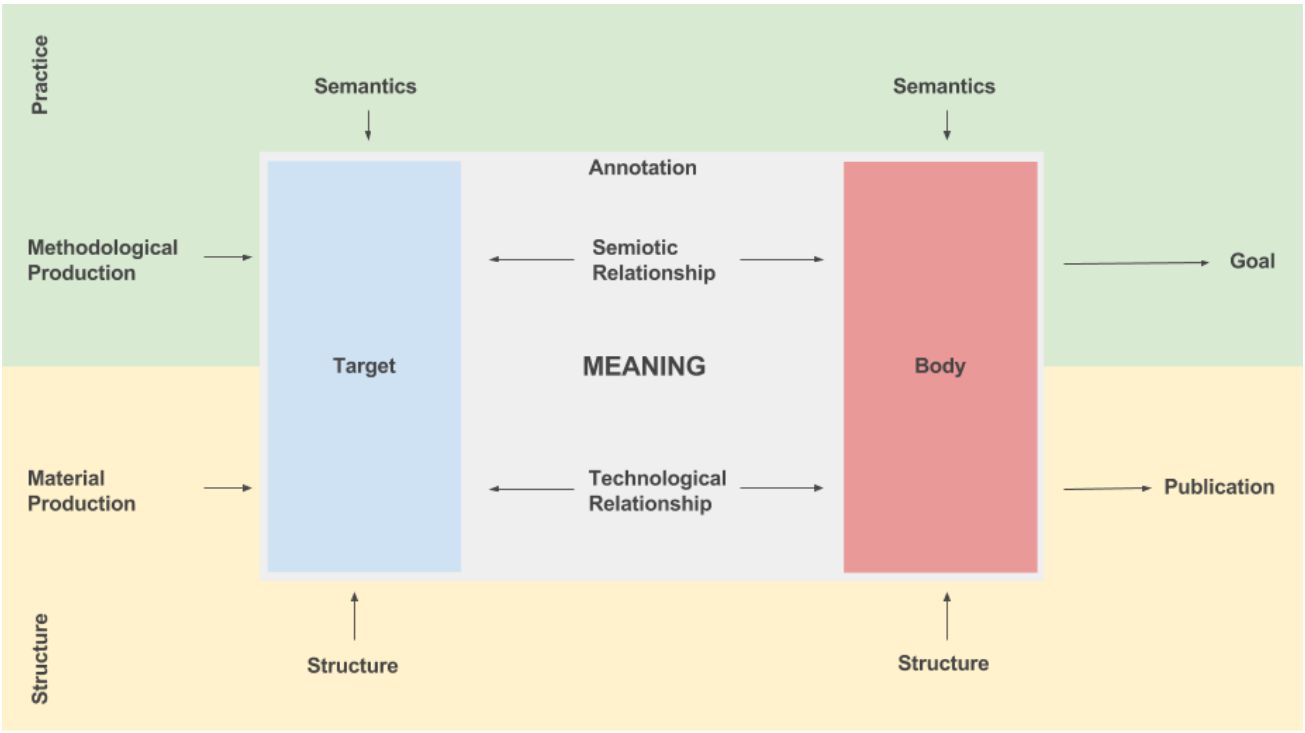

Figure 1: Meaningful context dimensions of annotations and annotated objects.

In other situations in which no technical representation of a formal vocabulary exists the formalization needs to be achieved by using other strategies. Dimensions like publication, goal, methodological or material production of annotations and annotation targets are even more complicated. Standards like Open Annotation are only partially supporting these dimensions albeit their relevance became clear in this section.

3. The Meaning of Annotations

I stated at the beginning of this paper that digital technologies have significantly multiplied the contexts in which annotations are used as a tool for research. Another way to look at this situation is to say that people speak of annotations in situations where th term would not have been applied before.

A productive way to look at this issue implies to conceive of this development as a co-dynamic. In this co-dynamic technologically defined annotation concepts and services are transfered to new research situations, modify the perception and concept of annotating on a theoretical level and are modified themselves by an updated discourse about annotations. For the purpose to illustrate these changes I want to give a few examples.

3.1. New Prospects in Contemporary Annotations

A traditional way to look at annotations implies a hierarchical relationship between the annotation and the object that is annotated. This relationship is visually well illustrated in medieval glosses which are often arranged around the text in the center of the page. The relationship exists also on the level of production. A book is produced for the main text and gives reason to add annotations 'in the margins'.

In definition of the concept of annotations within the Open Annotation Data Model this relationship vanishes. It says:

Annotating, the act of creating associations between distinct pieces of information, […] (Sanderson et al., 2013, p. 1)

The formal semantics still contain the concept of body and target but methodologically there is no necessary difference in the way Open Annotation understands the relationship between body and target.

In the Pelagios[15] project for instance sources are annotated with data about places. However, the services Pelagios provide completely blur this dependency. If the places are annotations to the texts or the other way around depends on one's point of interest.

This has to do with another principle of historical annotations that becomes more and more fragile: the existential dependency of annotations from the annotated object. Annotations exist on the paper of books and vanishes away with it. Digital annotations can be physically stored and disseminated independently. That means annotation data is a primary research output in itself and not solely anymore a documentation of the path that was taken to these results. By using Pelagios, annotations are brought to the level of first class research results.

Roorda (2013) addresses this aspect more explicitly when he calls annotations 'a new paradigm in Archiving'. In his opinion annotations are most importantly the smallest publishable information unit today. Such a use-case which completely abstracts from the linking aspect of annotations and highlights the information structure aspect

15 http://commons.pelagios.org/

is also elaborated in the Wissensspeicher[16] project of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW, 2016).

Roorda introduces another interesting prospect. In his Shebanq project (Roorda, 2016) he uses annotations for queries. In this scenario the annotated object is the volatile entity. The target of the annotation might look differently each time the annotation is rendered. The stable unit is the query as an interface, that means the representation of a certain way to look at things.

A better known issue is the shift of annotations from individual private contexts into open and collaborative spaces (Meister, 2012).

3.2. What is Annotating Today

All the examples demonstrate how fundamentally the scope of annotations and the deployment of annotating have changed by virtue of digital technologies. The level of change might be even more obvious considering that some of the main features of web annotations are just a real world implementation of how hypermedia research has envisioned the world wide web ever since (Carr, De Roure, Hall, & Hill, 1995; van Ossenbruggen, Hardman, & Rutledge, 2006). Measured by the quantity of researchers from hypermedia research which are active in the subject of web annotations this subject is hypermedia research. Thus, a technology and architecture oriented angle is driving our understanding of annotations.

The second question from the beginning of this paper asked: What are annotations today? The motivation is not to be essentialist or restrictive. The issue is that both the sound use of annotating in research and of annotation data as research resources depend on and improve with our understanding about digital annotations.. It is simply a question of methodology. The Digital Humanities community should therefor extend their otherwise often used concept of tools to discourse by definition. The benefit of such an approach is the creation of a deeper level of understanding of what is going on and consequently a more productive use of annotations. Definitions can be changed as Digital Humanists change their tools. It is the struggle which creates the benefit.

Fortunately, at least two recent initiatives aim at evaluating research practices in the Digital Humanities. The first is the Scholarly Domain Model (SDM) which appeared in the EUROPEANA project cluster (Gradmann et al., 2015) and was first defined in the DM2E satellite project[17]. The second initiative comes from the Digital Curation Unit in Athens[18] and started in the NeDiMAH project[19]. Its name is the 'NeDiMAH Method Ontology' (Digital Curation Unit, 2016; Hughes, Constantopoulos, & Dallas, 2016). In DARIAH-EU the DiMPO[20] is trying to use the NeDiMAH Method Ontology (hereinafter NeMO) to make progress on mapping Digital Humanities activities.

While NeMO's strategy adheres to a bottom-up strategy SDM at least partially partially follows a top-down approach. Nonetheless, both are well suited to record, structure and synthesize information about annotating today. Furthermore, both project clusters represent big communities which offer potentially rich content in this respect. In the case of SDM the primer even illustrates the model on the basis of an annotation example.

4. Feasibility, Strategies and Prospects

In this paper I demonstrated that the meaning which is embedded in annotations is fragile and often hard to grasp. I introduced a systematic approach to gain more control over the expression and interpretation of meaning in annotations. In the first case the systematology offers new insights for the application of metadata to annotations. I will come back to this issue below. In the second case it is a tool which can be used to look at or research on the context of annotation data before it is used. That being said, the systematology is still a device of understanding even if annotation data does not provide sufficient metadata.

I also indicated how fundamentally the concept of annotations is changing due to computational environments and argued in favour of broad evaluation of annotation activities. These two topics are only two different topics in the first place. In the long run, they belong to the same effort and contribute to each other. More precisely, the systematization of annotation activities into profiles will greatly enhance the understanding of context in annotation data. It will make it easier and more standard compliant to instantiate and describe context dimensions of annotation data. Likewise, deeper elaboration of the context dimensions systematology will make it easier to map annotation activities and identify profiles.

In the first section I indicated the complexity of annotation data and its reasons. The second section tried to make implicit things explicit and created a formal systematology. Furthermore, I criticized that current standardized annotation models like Open Annotation and its concepts of motivation and provenance are not sufficient. Thus, I am indeed arguing that annotation data needs more metadata applied to it than it is the case today.

However, it is also not feasible to describe annotation data in all its facets. The example of Semantic Web compliant tagging demonstrates that a complete description is not always necessary. The vocabulary in use is referenced implicitly in the namespace of the tag. In contrast, the e-Poetics use-case showed that these implicit means do not hold where no Semantic Web compliant vocabulary exists.

The information density of metadata which needs to be attached explicitly depends very much on aspects like the one that has just been described. It also depends on the structural and semantic model which is used to model annotation data as well as on the conditions for its instantiation in peculiar annotation scenarios. These scenarios create options to go without extra metadata or eliminate these options. Further research is needed to clarify which options for each context dimension exist in relation with which technological environments.

The argument of implicitly given informations can be pushed even further and up to the socio-cultural level. For

16 http://wissensspeicher.bbaw.de/
17 http://dm2e.eu/
18 http://www.dcu.gr/
19 http://nedimah.eu/
20 https://dariahre.hypotheses.org/working-groups/digital-methods-practices-and-ontologies

instance, the use of annotations for data publication in Bioinformatics is a straight-forward and well known practice. This means many informations about context dimensions have become part of common knowledge. In general, the level up to which this tacit knowledge exists for specific annotation scenarios influences the need for explicit metadata.

Computer science distinguishes between technological, structural and semantic interoperability. There is also something like socio-cultural interoperability which refers to questions of how public and consistent things are within a socio-cultural configuration. On the other hand this level of interoperability only exists insofar it is actively designed. In this sense the approach that has been presented in this paper tried to create better conditions for socio-cultural interoperability in annotating. Its success depends on further theoretical systematization of annotation activities in similar environments like DARIAH.

5. Disclaimer

The current study was financed by the Federal Ministry of Education and Research (BMBF) and carried out in the context of Cluster 6 in DARIAH-DE and the DARIAH-EU Working Group Digital Annotation. Special thanks go out to all projects that took the time to provide descriptions of their annotation use-cases.

6. References

Maristella Agosti, Giorge Bonfiglio-Dosio, Nicola Ferro. 2007. A historical and contemporary study on annotations to derive key features for systems design. In: *International Journal on Digital Libraries*, 8(1), pages 1–19. http://doi.org/10.1007/s00799-007-0010-0

Daniel Atkins. 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*.

Matthias Bauer, Angelika Zirker. 2015. Whipping Boys Explained: Literary Annotation and Digital Humanities. In: *Literary Studies in the Digital Age: An Evolving Anthology*. MLA Commons. https://dlsanthology.commons.mla.org/whipping-boys-explained-literary-annotation-and-digital-humanities/

Marie-Eve Bélanger. 2010. *Annotations and the Digital Humanities Research Cycle: Implications for Personal Information Management*. http://hdl.handle.net/2142/15035

Les Carr, David De Roure, Wendy Hall, Gary Hill. 1995. *The Distributed Link Service: A Tool for Publishers, Authors and Readers*. http://eprints.soton.ac.uk/250739/.

Chia-Ning Chiang. 2010. *A multi-dimensional approach to the study of online annotation* Ph.D thesis.

Stefan Gradmann, Steffen Hennicke, Gerold Tschumpel, Kristin Dill, Klaus Thoden, Alois Pichler, Christian Morbidoni. 2015. *Beyond Infrastructure! Modelling the Scholarly Domain*.

Lorna Hughes, Panos Constantopoulos, Costis Dallas. 2016. Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In: S. Schreibman, R. Siemens, J. Unsworth, eds., *A New Companion to Digital Humanities*, pages 150–170. John Wiley & Sons, Chichester.

Jan Christoph Meister. 2012. Crowd Sourcing 'True Meaning': A Collaborative Markup Approach to Textual Interpretation. In: M. Deegan, W. McCarty, eds., *Collaborative Research in the Digital Humanities*, pages 105–122. Ashgate, London.

Jan Chrsitoph Meister. 2015. *All dressed up and nowhere to go? The strategic role of digital humanities annotation tools*. Presentation. Hamburg. https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18469

Geoffrey Rockwell. 2010. As Transparent as Infrastructure: On the research of cyberinfrastructure in the humanities. In: M. Jerome, eds., *Online Humanities Scholarship: The Shape of Things to Come*, pages 461–487. Rice University Press , Houston.

Dirk Roorda, Charles Heuvel. 2013. Annotation as a new paradigm. In: *Proceedings of the American Society for Information Science and Technology*, 49(1), pages 1-10. Presentation. Baltimore. http://doi.org/10.1002/meet.14504901084

Robert Sanderson, Paolo Ciccarese, Herbert Van de Sompel. 2013. *Designing the W3C Open Annotation Data Model*.

Jacco van Ossenbruggen, Lynda Hardman, Lloyd Rutledge. 2006. Hypermedia and the semantic web: A research agenda. In: *Journal of Digital Information*, 3(1).

Niels-Oliver Walkowski. 2015. *Provenance and Motivation in Open Annotation*. Presentation. Cologne. http://cutuchiqueno.webfactional.com/slides/koeln052015

Niels-Oliver Walkowski. 2016. *Digitale Annotationen: "Best Practices" und Potentiale I,* Report No. 6.2.1 I. Göttingen, DARIAH-DE.