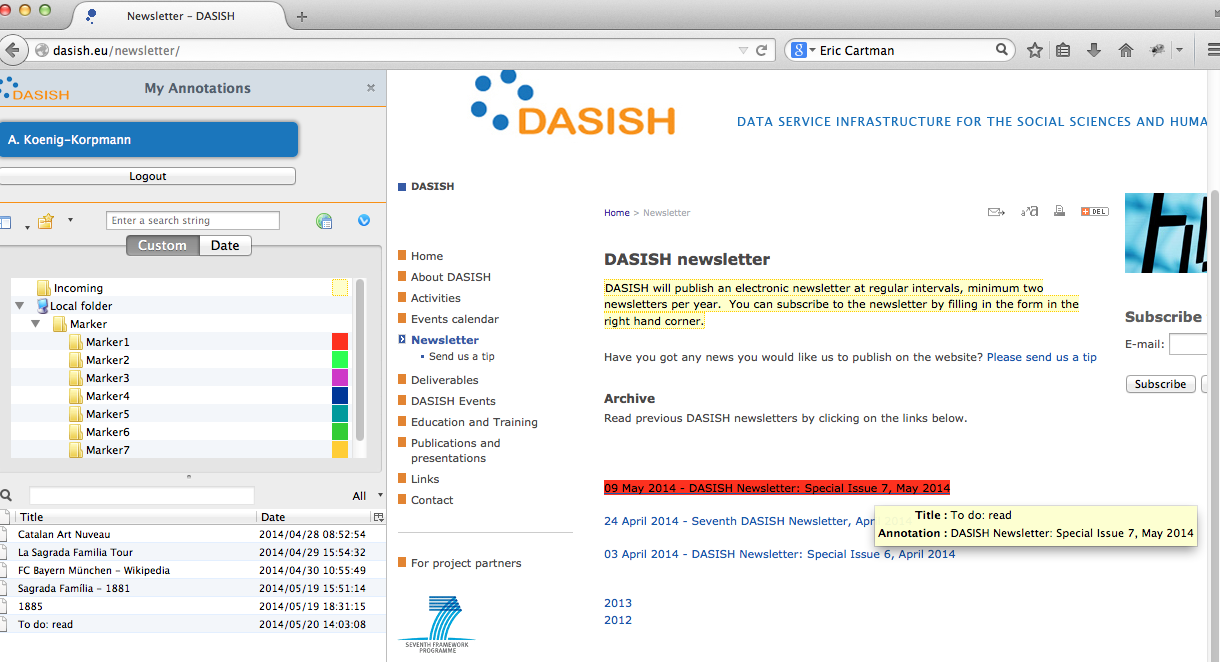
DASISH Web Annotator

Over the past decades, large amounts of data have been moved from paper and tapes to digital archives. Many of these archives are available online for the wide public or specific research communities. This allows multiple researchers to work remotely on the same data source, for instance, by annotating (making remarks over) the source and sharing these annotations within the group or publicly. Usually one speaks of annotations when a given content is processed and augmented by someone who is not necessarily the owner of this content and the action is taking place in another location than the one where the content resides.



There has been a good amount of tools for collaborative annotation. In the frame of the EU funded project DASISH (<http://dasish.eu>), The Language Archive at the Max Planck Institute for Psychonlinguistics (TLA-MPI) in collaboration with University of Gothenburg has been developing an annotation tool, DWAN, as well. DWAN is an abbreviation for DASISH Web Annotator. Its distinguishing feature lies in the fact that annotated sources, which may include regularly updated webpages, are stored (“cached”) in the database of a digital archive. This contrasts to storing only corresponding links or storing copies locally on the researcher’s computer. Archiving cached copies of annotated sources in the central database guarantees sustainability of shared annotations. The digital storage for annotations and related resources are provided by TLA.

To be more precise, DWAN is not a single tool but rather a framework for software annotation clients working together with a single back-end consisting of a database and a Representational State Transfer (REST) web service implemented in Java. Types of annotatable sources are defined by a client. The framework allows for the annotation of any web-accessible content, linking data, creating relations, or providing feedback. TLA has been developing the back-end and maintaining the database, whereas the colleagues from the University of Gothenburg significantly refactored an existing tool called “Wired Marker” (https://addons.mozilla.org/nl/firefox/addon/wired-marker/) to make it into a suitable DWAN client. “Wired Marker” is used to annotate webpages. However, within the DWAN framework one can annotate not only webpages in the HTML format, but also other document types such as XML documents generated by linguistic software. Like for example the EAF (MPI, 2010) file format used by the ELAN multimedia annotation software developed at TLA-MPI. TLA is developing the prototype of a specific DWAN-client for ELAN. The first results look quite promising. As for possible future work we mention the possibilty to annotate metadata records accessible by tools like the Virtual Language Observatory and CMDI Browser (both developed in the context of the CLARIN Component Metadata Infrastructure).

To be able to talk, i.e. obtain or submit annotations and the relevant information, to the database, a client must use a “common language” with the server maintaining access to the database. This common language comprises first of all the collection of REST requests recognizable by the server and listed in the documentation. To call one of the server’s REST methods, the client submits a request to a specified URL. Second, since a REST request is typically sent with its body, where the “content” is given in the form of XML structure, any such body must satisfy a certain structure, that is have a certain type corresponding to this request. The admissible types are given in DWAN’s XML schema. The schema formalizes DWAN data model, which strives to be compliant with the Open Annotation Data Model and Ontology developed by the Open Annotation Collaboration(The Open Annotation Collaboration and the Board of Trustees of the University of Illinois, 2014). Thus, the *Annotation* class is the core of the DWAN data model with the key relations, *Annotation* – *Body*, *Annotation* – *Target*, *Target* – *Source*, and *Target* – *Cached* *Representation*. These are used to define relationships between (1) annotation and its actual content, (2) annotation and the resource that is being annotated (i.e., the target), (3) the resource and its source URI, and (4) the target and a centrally stored structural or visual representation of this target. An instance of class *Annotation* is a structure containing essential information about a user-created annotation, such as annotation identifier, owner reference and time of creation.

A relational database provides storage for all annotations and related resources. A resource is stored in one of the five main database tables, in accordance with the resource type: annotation, target, cached representation, principal or notebook.

DWAN framework is to be presented at the forthcoming 9th Edition of its Language Resources and Evaluation Conference (LREC, <http://lrec2014.lrec-conf.org>) in Reykjavik.