



The goal of Data Service Infrastructure for the Social Sciences and Humanities (DASISH) is to determine areas of cross-fertilization and synergy in the infrastructure development and to work on concrete joint activities related to data, such as data access, data sharing, data quality, data archiving and legal and ethical aspects. As stated in DASISH task 5.6 description, researchers need to be able to store the results of collaborative intellectual work either as an annotation of a single fragment or in the form of typed relations between a number of fragments.

## DWAN: DASISH Web Annotator

### Using web-annotation tools

The use of the world-wide-web and the availability of on-line digital archives and other research data via the Internet creates new opportunities for research collaborations. Indeed, equipped with special software, researchers from different institutions and disciplines can work together via the network commenting existing content. Such collaboration can take the form of annotating data and sharing these annotations via an annotation sharing infrastructure. Such an annotation is a comment or other additional information pertaining to one or more fragments of a on-line document or set of documents.

### DWAN framework

DWAN is a software framework that facilitates client tools working together with a single back-end that

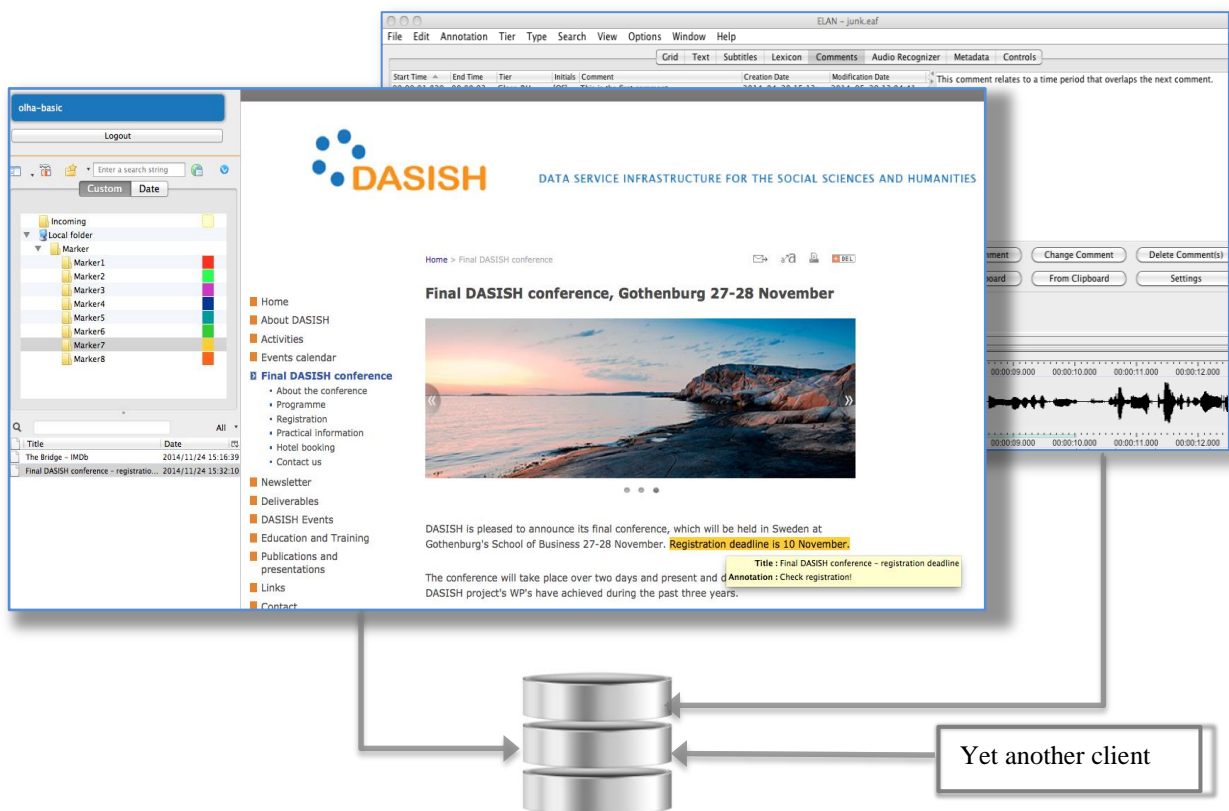
consists of a database and a server side application enabling data exchange between the remote clients and the database.

Clients can be developed for any particular sort of discipline specific data object, whereas the back-end is not domain specific and communicates with all clients in a generic way.

The back-end and its clients exchange data by sending HTTP-requests and processing responses. Requests and responses usually contain XML formatted messages that obey the common DWAN XML schema. This schema represents a data model that has been designed to represent the main data structures, which are involved in constructing annotations, and relations between these structures.

### DWAN data model

The relations *Annotation - Target*, *Target - Source*, *Target - Cached Representation* closely follow the *Open Annotation (OA)* standard<sup>1</sup>. An annotation, i.e. an



instance of the class *Annotation*, is a structure that contains essential information about an individual annotation. In particular it contains the annotation's identifier, the reference to the owner and the time of creation. An owner is either the user who has created the annotation or a user to whom the ownership has been assigned.

Besides the owner, an annotation has *readers* and *writers*. As one can expect, a reader is a user that can read the annotation, and a writer can also add changes to it. A user with *all* access rights for the annotation can delete the annotation or change its access rights for other users. Thus, a registered user can be related to an annotation by means of one of four access modes: *all*, *write*, *read* or *none*.

An annotation can have one or more *targets*. A target (i.e. an instance of the *Target* class) contains the reference to the web-document (a *source*) and the precise identification of the document's fragment that is actually annotated. Moreover, a target may refer to one or more *cached representations* of (the relevant parts of) the target document with the precise descriptions of the annotated fragments for each representation.

The semantics of an annotation is given in its body. In the implementation a body is an arbitrary text or an XML text.

## DWAN back-end

The core of the back-end is a database where annotations and information about annotated targets are stored together with *cached representations* of the targets. A cached representation is a copy or projection, e.g. a screenshot, of a target document or a fragment thereof. Storing cached representations enables the retrieval of a copy of an annotated document in case the actual web-document under the target's URI has been updated or removed, and locating the annotation in it becomes difficult or even impossible.

## Existing DWAN clients

**DWAN browser extension:** This client has been developed on basis of the freeware *Wired-Marker* extension<sup>2</sup>. It is a highlighter that allows marking fragments of a web-document by different colours. An annotated fragment can be a text fragment or the whole image. An annotated fragment is preserved not only in the local client's database within the extension but also sent by the DWAN client as an XML request to the back-end, where it is stored in the database together

with a cached representation of the annotated page (on the moment of annotation). A cached representation can be requested by the user later, for instance if the client cannot deliver the annotation because the page has been changed and the fragment cannot be resolved.

**ELAN client:** ELAN<sup>3</sup> is an annotation program for media files. An integrated client for the DWAN back-end is a result of the CLARIN NL ColTime<sup>4</sup> project. *Comments* in ELAN map to the concept of Annotations in the DWAN model. There is no specific support for comments on the ELAN annotations themselves. The DWAN backend is an ideal vehicle to store these comments: it is based on associating such comments with a URL.

**ANNEX client:** ANNEX is an open-source online visualizer for time-aligned annotation files, primarily targeted at the EAF (ELAN Annotation Format) format just as ELAN. Primarily being a visualization tool for archived materials, ANNEX currently does not offer any creation or modification functionalities over time-aligned annotation files. This leverages the need for extra functionality allowing users to create comments on the existing annotations.

For this purpose the DWAN back-end stands out as an ideal server-side engine to store, search and retrieve such comments.

## Outlook

Because of the universality of the DWAN data model and back-end implementation, the whole framework has the potential of being reusable by many research groups and institutions. An interested organisation can either reuse DWAN's backend code, or share the database with another organization, and create its standalone frontend, or adjust an already existing one. Amongst potential candidates we identified to be used as a DWAN frontend we would propose the following software: *ANNIS* (specific for linguistic corpora), *BIBLIOPEDIA* (designed to crawl scholarly resources), *LitBlitz* or *NB* (for faculties and students) *MapHub* (specific on annotating historic maps) *Pliny* (developed by KCL), *Pundit*, *Uvic Image Markup Tool* (specific for images), *Virtual Lighbox* (for Museums and Archives), *Zotero* (widely used citation manager), *Weblicht*.

## Contact

tla@mpi.nl

<sup>1</sup> <http://www.openannotation.org/>

<sup>2</sup> <http://www.wired-marker.org/>

<sup>3</sup> <http://tla.mpi.nl/tools/elan>

<sup>4</sup> <http://www.ru.nl/sign-lang/projects/coltime/>

