The DWAN framework: Application of a web annotation framework for the general humanities to the domain of language resources

Przemyslaw Lenkiewicz[1], Olha Shkaravska[1], Twan Goosen[2], Menzo Windhouwer[3], Daan Broeder[1], Stephanie Roth[4], Olof Olsson[4]

[1] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[2] CLARIN ERIC, Trans 10, 3512 JK Utrecht, The Netherlands
[3] DANS, Anna van Saksenlaan 51, 2593 HW The Hague, The Netherlands
[4] University of Gothenburg, PO Box 100, SE-405 30 Gothenburg, Sweden
{Przemek.Lenkiewicz, Olha.Shkaravska, Daan.Broeder}@mpi.nl,
twan@clarin.eu, Menzo.Windhouwer@dans.knaw.nl,
{Stephanie.Roth, Olof.Olsson}@snd.gu.se

Abstract

Researchers share large amounts of digital resources, which offer new chances for cooperation. Collaborative annotation systems are meant to support this. Often these systems are targeted at a specific task or domain, e.g., annotation of a corpus. The DWAN framework for web annotation is generic and can support a wide range of tasks and domains. A key feature of the framework is its support for caching representations of the annotated resource. This allows showing the context of the annotation even if the resource has changed or has been removed. The paper describes the design and implementation of the framework. Use cases provided by researchers are well in line with the key characteristics of the DWAN annotation framework.

Keywords: annotation framework, collaboration, semantic interrelations

1. Introduction

In the last decades, we have witnessed large amounts of data moving from researchers' drawers to digital archives. These archives have been connected to the Internet, spreading the content through the research community. The availability of such data presents new chances for collaboration. To bring this collaborative environment to a next, higher level, the requirement is to develop a set of tools that allows groups of researchers from different institutions, countries, or backgrounds to work together. Such collaboration can take the form of annotating the data, and sharing these annotations using an annotation infrastructure.

We typically speak of annotation, when a given content is processed, augmented or enhanced 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.

Collaborative annotation expresses the concept that an arbitrary number of people will be able to share the same annotations and can annotate the same documents or media files together, thus collaboratively making use of the interoperability features the annotation tool provides. A first step towards this goal is having the annotation data stored in a shared database.

There has, in fact, been a good amount of work in the domain of collaborative annotation. Some of the developed systems are aimed at specific tasks and/or domains. Examples of this are the Brat rapid annotation tool (Stenetorp, Pyysalo, Topić, Ohta, Ananiadou, & Tsujii, 2012) and GATE Teamware (GATE project team, 2014), which are especially aimed at annotating a textual corpus for NLP tasks. Other tools focus on the task of semantic annotation, e.g., Pundit (Grassi, Morbidoni, Nucci, Fonda, & Di Donato, 2013) and Semantic Turkey (Fiorelli, Pazienza, & Stellato, 2013). The Open Annotation Collaboration has developed a data model (Sanderson, Ciccarese, & Van de Sompel, 2013), which strives to become a W3C recommendation and enable easy exchange of annotations.

In this paper, we present the DASISH[1] Web Annotation (DWAN) framework, which is our proposal for the collaborative annotation solution. Its distinguishing features include a free form annotation body, i.e., adaptable to any task at hand, and special emphasis on supporting the dynamic nature of web resources, i.e., by allowing to cache (past) representations of a resource.

2. The DWAN framework

DWAN is 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. It allows annotating any web-accessible content, linking data, creating relations, or providing feedback. Its novelty is also in the fact that the created content and sources can be stored in a digital archive, which guarantees their sustainability and persistence. The digital storage for annotations and related resources is provided by TLA-MPI[2].

DWAN is also especially meant to cater for specific linguistic tools that through their use of linguistic data formats can annotate specific linguistic items such as lexical items, annotation tags etc.

2.1 Specification

Our first step was to define a data model capable of processing free-text annotations consisting of notes, descriptions, commentaries, and critical examinations of

[1] http://dasish.eu/

[2] The Language Archive, Max Planck Institute for Psycholinguistics, http://tla.mpi.nl/

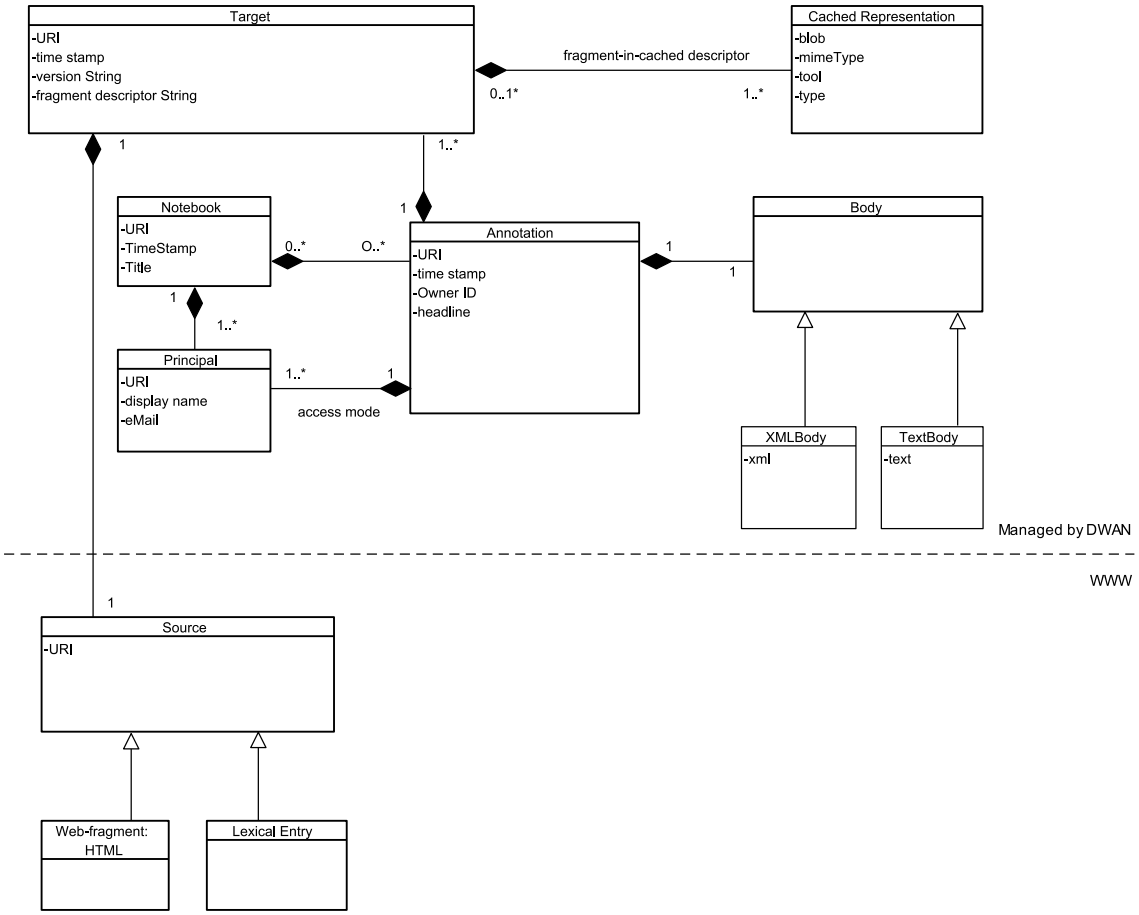

Figure 1: The DWAN data model

fragments of web-accessible documents. The latter include both web pages in HTML format and other document types such as XML documents generated by linguistic software. One such example is the EAF (MPI, 2010) file format created by the ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) multimedia annotation software developed at TLA-MPI. But also metadata records as made accessible by tools like the Virtual Language Observatory and CMDI Browser (both developed in the context of the CLARIN Component Metadata Infrastructure (Broeder, et al., 2010)), can be annotated. The object-oriented implementation of DWAN enables us to adapt the framework to future use with linguistic [3] and corpora annotations[4].

The DWAN data model, presented in Figure 1, strives to be compliant with the Open Annotation Data Model (Sanderson, Ciccarese, & Van de Sompel, 2013) 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. In this paper and the DWAN specification, a 'user' is a principal[5], which may represent either a single user or a group of users. An annotation can also have *readers* and *writers*, who are allowed respectively to only view an annotation or to change it.

An instance of *Annotation* has one or more target relations. An instantiation of the *Target* class contains a

[3] Linguistic annotations are definitions of linguistic features (regarding e.g. grammar, semantics, syntax, morphology) of the annotated text giving information about the words and sentences of the text.

[4] Annotated corpora can serve as repositories of linguistic information containing explicit information through specific annotations.

[5] By definition, a principal is represented by an account, a role, or some other sort of unique identifier. Principals are the unique keys used in access control lists. They can represent human users but also identification codes for automated connection access via applications. (Stackoverflow, 2014)

reference to the web-accessible document (i.e. a *Source* object) as well as a precise pointer to the annotated document fragment. Moreover, a target may refer to one or more cached representations of the relevant parts of the source with detailed descriptions of the annotated fragments for each representation.
The actual content of an annotation is included in its body. The framework data model specifies the body content to be of type text or XML. However, the exact structure of the text or XML content is not specified, which enables arbitrary annotation bodies as long as the information can be serialized as text or XML.
Finally, for ease of sharing and user collaboration, annotations can be organized in so-called notebooks with specific access permissions and restrictions.
This data model forms the core of the DWAN framework (see Figure 2), i.e., the central database and the REST API are based on this model.

2.2 Database

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.
The encapsulating service ensures the right sequencing of editing and deleting operations. For instance, an annotation is not deleted before all the records of the related targets are removed.
The DWAN framework allows storing a cached copy for each version of the target resource, whenever an annotation is performed. When the target resource changes it is possible, client-side, to either see the cached copy that was annotated, or try to 'remap' the annotation to the updated resource, which is done by content matching.
Keeping all the versions of the created annotations, as well as the target resources, is a very important and unique feature of DWAN. Other available tools, like Pundit or Memento (Van de Sompel, Nelson, Snaderson, Balakireva, Ainsworth, & Shankar, 2009) rely on web masters and archives to provide previous versions of resources.

2.3 REST Interface

A client accesses the annotations by means of methods provided by a REST interface available over HTTP. To call one of the server's REST methods, the client submits a request to a specific URL. To access an existing annotation, the client needs to send the annotation's external identifier, which is automatically generated by the server when the annotation is added to the database. The service also provides methods to request all annotations on a resource accessible for a specific user. Client-server communication is exclusively handled with REST requests.
The current DWAN implementation uses a traditional, but well-known and stable, software stack, i.e., a relational database for persistent storage and an object oriented application server supporting the application programming interface (API) for producing the XML representations. The following example shows a DWAN resource XML representation[6].

```
<annotation
  xmlns="http://www.dasish.eu/ns/addit"
  ownerRef="dwan:/api/users/111"
  URI="dwan:/api/annotations/2c">
 <headline>Wichita example</headline>
 <lastModified>2013-08-12T11:25:00.383Z</lastModified>
 <body>
  <textBody>
   <mimeType>text/x-markdown</mimeType>
   <body>Found a *nice*
     [annotated video recording]
     (dwan:/api/targets/3c) of
     [Wichita](dwan:/api/targets/32)
     speakers!</body>
  </textBody>
 </body>
 <targets>
  <targetInfo
    ref="dwan:/api/targets/3c">
   <link>hdl:1839/00-0000-0000-0017-E55B-3</link>
   <version>2012-10-30T16:18:09.000Z</version>
  </targetInfo>
  <targetInfo
    ref="dwan:/api/targets/32">
   <link>http://wals.info/languoid/lect/wals_code_wic</link>
   <version>96160aefa8fba01d26e140ae6138c… </version>
  </targetInfo>
 </targets>
 <permissions>
  <userWithPermission
    ref="dwan:/api/users/111">
   <permission>reader</permission>
  </userWithPermission>
  <userWithPermission
    ref="dwan:/api/users/112">
   <permission>writer</permission>
  </userWithPermission>
 </permissions>
</annotation>
```

However, the REST API is flexible and allows, in particular, for the addition of means of representation for transport, e.g., Open Annotation RDF or JSON, via HTTP content negotiation.

2.4 Implementation of a browser-based web annotation client

For the first prototype implementation of the web browser-based DWAN client, we chose to use and extend the Wired-Marker [7] add-on. This Firefox extension

[6] Due to space considerations the development URLs have been shortened by using a dwan: prefix.
[7] http://www.wired-marker.org/en/

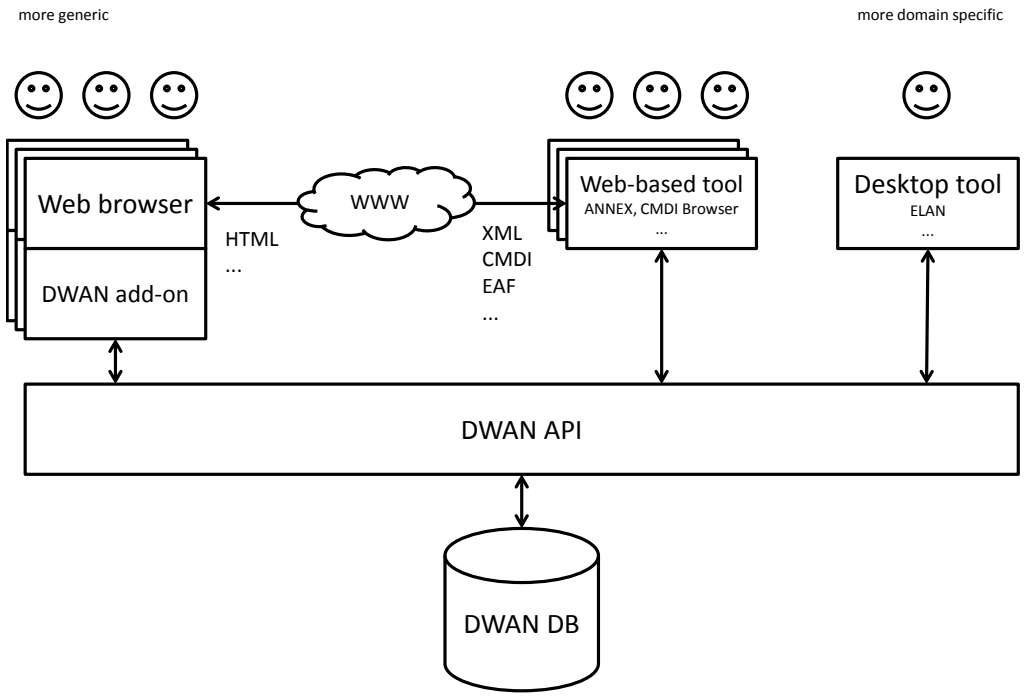

Figure 2: The DWAN Framework

enables the user to create free-text annotations on fragments of webpage content. The annotations can be assigned to *categories* that are visualized as a tree of *local folders* and assigned a style. The user-produced data comprising annotations, annotation titles and webpage titles are stored in a local database. The realization of extended functions for the DWAN client was focused on creating collaborative annotations by implementing communication with a central server.
Moreover, the DWAN client contains upgrades concerning graphical components, deactivation of redundant modules, branding, and compatibility fixes.

2.5 Other clients

In the DWAN framework, we allow the usage of any client application that can connect to the back end using the defined REST API. This creates the possibility to annotate any domain-specific content in a suitable way, as the information about exact properties and structure of the annotated content can be transmitted between the client and the annotation framework.
Furthermore, tools that already exist and are widely used by researchers can be extended with DWAN support and allow for annotation of any content that would be impossible to properly annotate otherwise. Also, the visualization of annotations can be performed in a better way than through a generic client, as the clients can have direct access to the annotated data. This is not possible e.g. in case of objects rendered in web pages, as the underlying data is typically not directly available.
Examples of such domain-specific tools that can be extended with annotation features can include lexical tools (e.g. LEXUS[8]), which would enhance collaboration on creation of lexica and linking lexical entries with examples or descriptions that can be found on the web, linking data to an ontology or taxonomy or another kind of semantic registry.

3. Use cases for language resources

We have performed a series of interviews with linguistic researchers at MPI, Språkbanken[9] and SND[10]. The aim was to confirm the original use cases provided by the DASISH community, which triggered the development of these kinds of annotation tools. Another aim was to look for additional possible use cases and suggestions for desired functionality. The desired properties for an annotation framework were focused on an adaptable, flexible tool that is not dependent on a certain linguistic theory or tradition and offers backup functionality for all annotation data via local or remote storage. These desires are well in line with the specification of the DWAN framework.
Several examples for use cases have been presented during these interviews. They are briefly presented in the sections below.

3.1 Use case: Free-text notes for literature research

In the domain of literature research, annotations are often used in the form of free-text notes, which are defined as descriptions, commentaries, and critical examinations of items in a publication. Today it is possible to create annotations electronically and link them with multiple data objects. Therefore, the researcher is able to achieve

[8] http://tla.mpi.nl/tools/tla-tools/lexus/

[9] http://spraakbanken.gu.se

[10] Swedish National Data Service, http://snd.gu.se

more than only noting down topics, examples, and pinpointing what is most relevant. It was stated that there is a great demand for a tool that is able to create cross-references, cross-relationships and relations between annotations, thus generating semantic interrelations and hierarchies. These relationships need to be established between fragments of the same source document or between fragments that are located on other remote data sources.

Furthermore, in the research community, there is a need for annotation instruments that can help with collaborative epistemic discourse with the possibility of private, in-group and public sharing of annotations. The interviewee knew no such comprehensive tool that could be applied for web content. Instead, several e-book reader applications with annotating functionality were frequently used. This emphasizes the gap between current requirements and framework implementations that have recently been released or are currently being done.

3.2 Use case: annotations for research in field linguistics

Cross-domain research in field linguistics (dialectology research; endangered or extinct languages research), language history and cultural anthropology requires linguistic annotations, which include morpheme analysis with the outcome of so-called morpheme-based grammatical annotations. The results of this type of field linguistic annotations will be saved in a database containing morphemes, allophones, and word roots. The benefit of generating this sort of data collections is that automatic annotation of e.g. other sound tracks will be possible. Moreover, lexica containing the new words, their definitions, and attributes can be built up. Even phonetic transcriptions of audio tracks will be annotated in the way specified above.

The interviewee had worked with the linguistic annotator ELAN previously, but knew of no tool that would allow a collaborative creation of linguistic annotations. Web-collaborations for sharing annotated language resources or to create linked open data would be a very interesting feature for linguistic or interdisciplinary transnational research collaborations. For example, in a project like DOBES (Wittenburg, Mosel, & Dwyer, Methods of Language Documentation in the DOBES project, 2002) (Himmelmann, 2006) teams could review each other's tiers in ELAN files in the form of annotations.

The need to take ethical issues (i.e. source and informant secrecy, ethnic traditions etc.) into account when implementing collaborative functionality and open data solutions was emphasized.

4. Conclusions

In this paper, we have presented the DWAN annotation framework, which instruments on-line, collaborative annotation of documents available via the Internet. Annotations can range from free-text notes to specialized linguistic markings with a fixed structure and semantic disambiguation.

In the proposed framework all annotations and related resources, including links to annotated sources and versioned copies thereof, are stored in a central database and handled by a central server. Different clients, covering different use cases, have access to the database via a uniform service interface.

The main distinguishing feature of the proposed framework is the possibility to cache representations of annotated resources. This allows for the preservation of the connection between the annotation's content and the corresponding version of the resource even after it has been significantly modified at its original location.

The design of the annotation framework was based on various use cases provided by the DASISH user community. To validate the DWAN framework we have performed a range of interviews with researchers from the field of literature and linguistics (see Section 3). Their use cases and desired features are well in line with the key characteristics of the DWAN annotation framework.

References

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., et al. (2010). A Data Category Registry- and Component-based Metadata Framework. *Seventh International Conference on Language Resources and Evaluation.* Malta: ELRA.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., & Di Donato, F. (2013). Pundit: Creating, Exploring and Consuming Semantic Annotations. *Proceedings of the 3rd International Workshop on Semantic Digital Archives.* Valletta, Malta.

Himmelmann, N. (2006). Language documentation: What is it and what is it good for. In J. Gippert, N. Himmelmann, & U. Mosel, *Essentials of language documentation.* Mouton - de Gruyter.

MPI. (2010, December). *ELAN Annotation Format*. Retrieved March 18, 2014 from http://www.mpi.nl/tools/elan/EAF_Annotation_Format.pdf

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013, February 8). *Open Annotation Data Model*. Retrieved March 18, 2014 from http://www.openannotation.org/spec/core/

Stackoverflow. (2014, January). *What is the meaning of Subject vs. User vs. Principal in a Security Context?* Retrieved March 18, 2014 from http://stackoverflow.com/questions/4989063/what-is-the-meaning-of-subject-vs-user-vs-principal-in-a-security-context

The Open Annotation Collaboration and the Board of Trustees of the University of Illinois. (2014, March 13). *Open Annotation Collaboration*. Retrieved March 18, 2014 from http://www.openannotation.org/

Van de Sompel, H., Nelson, M., Snaderson, R., Balakireva, L., Ainsworth, S., & Shankar, H. (2009). *Memento: Time Travel for the Web.* CoRR.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A.,

& Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559). Genoa, Italy: ELRA.

Wittenburg, P., Mosel, U., & Dwyer, A. (2002). Methods of Language Documentation in the DOBES project. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Canary Islands, Spain: ELRA.