

# The Documentalist Support System: a Web-Services based Tool for Semantic Annotation and Browsing

Hennie Brugman<sup>1</sup>, Véronique Malaisé<sup>2</sup>, Luit Gazendam<sup>3</sup>, and Guus Schreiber<sup>2</sup>

<sup>1</sup> MPI for Psycholinguistics, Nijmegen, The Netherlands

<sup>2</sup> Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands

<sup>3</sup> Telematica Instituut, Enschede, The Netherlands

**Abstract.** The Documentalist Support System (DocSS) is developed to suite novel needs of documentalists working within the Dutch archive for Sound and Vision, broadcasters working outside of Sound and Vision and people interested in the Cultural Heritage value of the archive, who want to perform search in context. The documentalists (and to some extent the other users mentioned) need an environment in which they can view and manipulate multiple types of information (documents and metadata), receive annotation suggestions from their controlled vocabularies, create catalogue descriptions and browse for semantically similar documents within their collection. The open architecture, the publicly published annotation format and the SKOS representation requirement for the controlled vocabularies make the generated annotations interoperable with other annotation databases and the DocSS usable for any documentalist annotating material with controlled vocabularies, for which a digital textual representation of the data exists.

## 1 Introduction

Annotation is one of the central processes in the management of archives: it anticipates on future usages of the archives and attaches information judged relevant for an optimal retrieval of the stored data. Most of the data which are to be archived come with a textual description, which can be used as a basis for generating annotation suggestions. Automatic annotation suggestions speed up and systematize this annotation process, which is nowadays still performed mostly manually. Annotations are usually meant for retrieving individual documents independently from one another, but an archive has also the power of showing documents in their socio-cultural context. This display in context can also be achieved based on the annotations, using the principles of Semantic Browsing. In the CHOICE project<sup>4</sup>, funded by the Dutch NWO<sup>5</sup>, we have developed a Web-Services based tool to generate ranked annotation suggestions and to navigate semantically through the documents of a corpus, representing a portion of

<sup>4</sup> <http://www.nwo.nl/CATCH/CHOICE>

<sup>5</sup> <http://www.nwo.nl/>

the archives. Although it was created for the purpose of one specific archive, the Netherlands Institute for Sound and Vision, it is a modular environment. The modularity is demonstrated by the implementation of three different use cases. The architecture is based on an annotation model formally defined according to Semantic Web standards, which makes the interaction and interoperability with the system easier. We present the semantic annotation pipeline that is used in the Documentalist Support System (DocSS) in section 2, the possibilities of semantic browsing in section 3, the implementation choices that were made to get a generic and modular architecture (section 4), and the three scenarios that are implemented in the current prototype (section 5). We conclude with evaluation ideas and general perspectives about this system.

## 2 Semantic Annotation and the CHOICE pipeline

Semantic annotation is the act of attaching metadata information about the semantic content of a document. This operation, often based on a controlled vocabulary, can range from manual [7] to automatic [1] process. In our project, we opted for an automatic generation of metadata, because our aim is to provide support to documentalist at annotation time: the documentalists would only select the relevant suggestions from the list. This configuration is quite common in real-life archives or Cultural Heritage institutions, and our tool can be applied to or reused in any case where the data to be annotated is different from the text that describes it: libraries, museums, etc.

We based our annotation on the GATE framework [2], a platform for NLP which is widely used in the community and implements basic functions (and a programming language for specifying linguistic rules) for Information Extraction. We co-developed Apolda [10], a plug-in that takes a thesaurus in the SKOS [8] format as input and annotates text segments with the URI's of the different matched concepts. SKOS is likely to become a W3C recommendation and has been used by different organizations to represent their controlled vocabularies, making this plug-in potentially interesting for many Cultural Heritage institutions in their annotation process.

The annotation is based on string matches, without contextual disambiguation rules in the corpus itself. We use the controlled vocabulary's structure as background knowledge to perform a ranking, which also has the property of performing disambiguation [9]. The implementation is therefore language independent and the ranking is based on frequency information and the vocabularies structure. This ranking is necessary because of the length of the suggestion lists: typically a list of 200 annotations is extracted from a text of 1500 words, making an unranked list useless as a proposition for annotation in a real-life process.

The first ranking that we implemented was CARROT [5]. CARROT takes into account the number of direct and indirect links that exist in the controlled vocabulary between the extracted terms. For example, the terms *War* and *Prisoners of war* are directly linked in the vocabulary used at Sound and Vision. The relations of the term *Prisoners of war* indirectly links the term *War* with *Pris-*

*ons*. The combination of direct and indirect relations creates graphs of extracted terms; CARROT groups the extracted terms into four groups of decreasing relevancy, depending on the local connectedness of a term in the graph. Subsequently each group is sorted on frequency or tf.idf. This ranking has proven to improve upon the frequency and tf.idf results.

We implemented the annotation and the (tf.idf and CARROT) ranking as Web Services, and integrated them in an interface where (textual) documents, besides being used as basis for generating annotation suggestions, can also be searched and browsed. Indeed, these documents constitute a rich context for the archived data, and can be used as an additional entry point for querying the archives where the data is stored. We describe the Semantic Browsing in more details in the next section. Note that the annotation web service can be applied both offline (where the resulting annotations are stored in a repository) and online (where the annotations are not stored but immediately used in the interface), while the ranking web service is always used dynamically.

### **3 Semantic Browsing**

Semantic Browsing is the act of navigating from document to document based on semantic features [3]. Semantic Browsing can be based on Semantic Annotations. These show all the possible semantic dimensions of a document, and constitute indirect links between the documents. A similarity measure for documents can be calculated and documents can be ranked on basis of the set of links between them. This functionality was implemented in a previous prototype, but has not been integrated in the DocSS yet.

Although our DocSS does not implement semantic browsing in the above sense, it creates a data set that is well suited for it. On the one hand, a digital resource (e.g. a radio or television program) is characterized by a ranked list of concepts that are derived from a set of text documents that were manually associated with this resource. On the other hand, each of the text documents is associated with a ranked list of concepts itself. Starting with some text document, we can then retrieve a ranked set of semantically similar digital resources. Or, alternatively, for some digital resource we can retrieve semantically similar resources. This type of Semantic Browsing functionality is less useful for documentalists than it is for people interested in the Cultural Heritage value of the researched archives.

### **4 Implementation choices for modularity and interoperability**

Although implemented in the context of a specific project, in which it satisfies the needs of one Audiovisual archive, the tool is implementing a publicly published annotation format (see section 4.1), which makes it interoperable with other annotation databases; as we mentioned earlier, its implementation makes

it language independent (section 4.2). Finally, it is an open architecture, enabling to load, process and open new documents, from which annotations can be generated based on the vocabulary that the user selects (see the third use case description, section 5.3).

#### 4.1 Public Annotation Model

The Semantic Annotations are based on the GATE API and Apolda plugin, which have an XML representation of the annotations. We developed, on top of this one, an annotation format formally defined in OWL [6], which is used in a set of research projects associated with Cultural Heritage-related institutions: the CATCH program, which includes CHOICE. The extensible model is expressed in RDF and contains core annotation properties that can be adapted to a particular media or research area. The strong point of the model is the possibility to anchor an annotation either to a document, to a part of a document, to an annotation or even to a part of an annotation, making it extremely flexible. With this model, different experts can contribute at different levels of annotation: an image featuring a handwritten document can be annotated with its represented textual content, parts of this content can be in turn annotated with metadata selected from a controlled vocabulary. For example, if the image of the handwritten text mentions Amsterdam, the string *Amsterdam* can be attached as a first level of annotation to the relevant image region. Upon this layer the URI of this place name from the Geonames database's RDF representation<sup>6</sup> can then be attached, adding a formal semantics to the first layer of informal annotation.

#### 4.2 Language Independent Implementation

As we have mentioned earlier, the Semantic Annotation is based on different lexical representations of a concept, and relies on (longest possible) string matching, it is language independent. The quality of the controlled vocabulary, and particularly the quality of the relationship's network between terms is the main guaranty for the quality of the ranking; this point is also language independent. We have tested the DocSS tool with two different languages, Dutch and English (see the Sound and Vision and the INCCA use cases). The extension to other languages (including ones with a non-roman script system) mainly relies on GATE's capacities. The method itself is generic enough to be applied to any language.

#### 4.3 Web service based architecture

Our basic design choice for developing DocSS was to implement it as a web application that realizes its semantic annotation and browsing user task support by means of orchestration of a number of web services. Several different web services are involved. One group of services is data oriented:

<sup>6</sup> See <http://www.geonames.org/ontology/>

- Text repository service. This service supports retrieval and uploading of text documents. Uploaded documents are automatically indexed using Lucene for retrieval on basis of full text search.
- Annotation repository service. This is used for storage and retrieval of several types of annotations that are all represented according to our annotation meta model. Annotation types involved in our current system are different varieties of object metadata (for radio/television programs, for associated text documents, for interviews with artists) as well as semantic annotations of text segments. The repository is implemented on basis of RDF (using Sesame) and our annotation repository web service API is implemented using Sesame’s query language, SeRQL.
- Vocabulary repository service. All vocabularies involved are represented using SKOS, and stored and accessed using a central repository. This repository will also support creation and maintenance of concept mappings. For access to both vocabularies and mappings a web service API is designed and implemented.

Another group of web services that take part in DocSS are services that encapsulate algorithms:

- Semantic annotation service. This service takes any text and annotates it using GATE and the Apolda plugin. The resulting annotations can be directly used or stored in the annotation repository for later retrieval. Apolda can be configured to use different vocabularies.
- Ranked recommendation service. Input for this service is a set of concept identifiers. The service reorders the list using one of several alternative ranking algorithms: TF.IDF, CARROT or our adaption of PageRank.

Note that both the annotation service and the ranking service make use of the vocabulary repository.

Using this modular architecture has as benefit that different components of the system can be distributed over several servers and/or sites. Since all components comply to clear and well-defined interfaces it is relatively easy to plug in new implementations or other components. The repositories used (annotations and vocabulary) can be shared by a community of users and are valuable resources by themselves. Finally, the services and repositories can be reused to implement other usage scenarios. On the downside, implementing an application as an orchestration of web services can create complex systems having lots of interdependencies.

## 5 Three use case scenarios

### 5.1 The Sound and Vision use case

Sound and Vision is the Dutch national audiovisual archives, which hosts a collection of 700,000 documents, this number is yearly increasing by 30.000 documents. In the CHOICE project, we manually created a corpus of 258 textual

resources related to TV programs archived at Sound and Vision: the main goal of the project is to propose annotation suggestions to the documentalists of Sound and Vision, and we could thus compare our propositions with existing manual annotations, by choosing our corpus within TV programs already archived and annotated at Sound and Vision. All the TV programs that we have selected for this test corpus are documentaries, as they usually have an extensive textual description, provided by the broadcasters themselves, and as other textual descriptions are still available even as long as 7 years after the broadcast. We considered that these generic descriptions of the TV programs would give us annotation extractions at a level of generality that would be relevant for good annotation suggestions.

The annotation suggestions are created with the Apolda plugin for GATE, and follow the annotation schema detailed above. In previous experiments [4], we have shown that we achieve 49% percent of precision at 10 and 64% of recall at 10 in best cases. This corpus and thesaurus are both in Dutch.

## 5.2 The INCCA use case

The second use case that we demonstrate is related to the INCCA project<sup>7</sup>: an international network of professionals connected to the conservation of modern and contemporary art and was established to meet the need for an international platform for knowledge and information exchange. The database that gathers the information from more than 100 organisations is the result of almost 10 years of work. The Netherlands Institute for Cultural Heritage, ICN<sup>8</sup>, is the coordinating institution of INCCA, and the one that manages the database.

In the course of the project, a list of controlled vocabulary has been gathered. One of the bottlenecks of INCCA is to annotate all of the information that is sent to the central database for sharing and reuse. We used the CHOICE pipeline to test whether we could generate relevant annotations for artists interviews based on the INCCA controlled vocabulary. We were given a set of 36 artists interviews and the controlled vocabulary, which was merely an alphabetical flatlist. Our first task was thus to create relationships between the terms, as our ranking algorithms are based on the background knowledge base's structure. We used NLP techniques to derive semantic relationships between the terms of the thesaurus themselves and automatically created an anchoring to WordNet to derive more relationships. These relationships were validated by an expert, and the generated annotations were evaluated by another expert. The annotations were here at a different level than what the expert evaluator would have chosen as annotations: he was annotating the underlying semantic aspects of the interview and the theoretical views of the artists, whereas the texts themselves contained mostly terms at a technical level: type of coating, painting techniques etc. As the humanly generated annotation's information is not present in the text, our method could obviously not extract it, but the annotations that we

---

<sup>7</sup> <http://www.incca.nl/Dir003/INCCA/CMT/Homepage.nsf>

<sup>8</sup> <http://www.icn.nl/>

generated are nevertheless useful: these annotations can help to answer low-level information needs that users of INCCA could have, and that would be too time consuming for an art expert to generate. Moreover, based on our annotations, we could relate artists that share sets of practices or materials, which is a novel way of investigating relations between artists, and might be interesting for a curator as an exhibition theme. Therefore, although different from human-made annotations, our automatic annotations provide still an interesting access point to the INCCA data. Both the corpus of interviews and the INCCA thesaurus (that we also converted to SKOS) are in English.

### 5.3 The Open Document use case

To show the modularity of the DocSS, a third use case was implemented: this scenario enables the user to upload a new (set of) document(s), and have it or them annotated by the terms of the thesaurus that is selected. The chosen textual resource(s) can be in any language supported by GATE, and the thesaurus can be selected between the GTAA, the INCCA thesaurus and a generic English thesaurus, also represented in SKOS: the UK Archival Thesaurus (UKAT)<sup>9</sup>. The UKAT is a subject thesaurus which has been created for the archive sector in the United Kingdom; its backbone is the UNESCO Thesaurus, a high-level thesaurus with terminology covering education, science, culture, the social and human sciences, information and communication, politics, law and economics. We uploaded the UKAT in our system for this demo, but in a further stage the user will also be able to define and upload the thesaurus or ontology of his choice, provided that the format is compatible with our tool.

## 6 Conclusion and perspectives

We successfully managed to build first versions of an extensible and pluggable web service based architecture and its components (both repositories and processing resources). We also built the DocSS web application combining these repositories and services into a support system for automatically suggesting keywords to documentalists who are archiving digital objects. We were able to show the system's generality, configurability and flexibility by implementing three different use cases, using different text corpora and background knowledge resources.

Semantic annotation and ranking are the main research topics of the CHOICE project. This software development work shows the feasibility of the researched application scenarios, it allows to easily inspect the resulting output of our algorithms and it allows us to test quality of recommendations and the impact on documentation practices with real documentalists.

In the process, we started building a rich graph connecting different heterogeneous archive resources and concepts from vocabularies. Text document sets

---

<sup>9</sup> <http://www.ukat.org.uk/>

can be associated with digital objects. These objects can have metadata that associates them with with concepts. The text documents are associated with concepts as well by means of semantic annotations. The thesauri we used contain links between concepts (both manually created and automatically added enrichments). This rich semantic graph offers interesting possibilities for semantic browsing and searching. Entry points for this are object metadata, thesaurus concept graphs or semantic annotation values (as well as free text search).

Next to improvement of the user interface and the quality of the annotation and ranking algorithms on basis of evaluations by professional documentalists, future work includes integration efforts in two different directions. First, the system (or the functionality that it offers) has to be integrated with the professional cataloging and workflow management system that operates at Sound and Vision. Second, in the CATCH program similar software is developed for generating annotations and recommendations for several other Cultural Heritage collections and media types (e.g. speech, musical melodies, video or scanned handwritten document images). DocSS and its services can be seen as a special case of this CATCH software development effort, and will be integrated with it.

Finally, on basis of the same set of web repositories and services other interesting web applications for non-documentalist users can easily be developed. We hope to experiment with scenarios for semantic document browsing, as well as with online involvement of broadcast archive users with the documentation/annotation process.

## Acknowledgments

This research was conducted in the NWO funded CHOICE project. We thank our colleagues, both at Sound and Vision and at our respective research institutes. A particular thank to ICN for providing us with their data and evaluating our results!

## References

1. Kiryakov A., Popov B., Terziev I., Manov D., and Ognyanoff D. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2(1):49–79, 2005.
2. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.
3. Alexander Faaborg and Carl Lagoze. Semantic browsing. In *ECDL*, pages 70–81, 2003.
4. Luit Gazendam, Veronique Malaise, Hennie Brugman, and Guus Schreiber. Comparing background-knowledge types for ranking automatically generated keywords. *16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW'08)*, 2008.

5. Luit Gazendam, Veronique Malaise, Guus Schreiber, and Hennie Brugman. Deriving semantic annotations of an audiovisual program from contextual texts. In *Proceedings of First International workshop on Semantic Web Annotations for Multimedia (SWAMM 2006)*, 2006.
6. Véronique Malaisé Hennie Brugman and Laura Hollink. A common multimedia annotation framework for cross linking cultural heritage digital collections. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.
7. J. Kahan and M.-R. Koivunen. Annotea: an open rdf infrastructure for shared web annotations. In *World Wide Web*, pages 623–632, 2001.
8. Alistair Miles and D. Brickley (editors). *SKOS Core Guide. W3C Public Working Draft*. World Wide Web Consortium, November 2005.
9. Hennie Brugmann Veronique Malais'e, Luit Gazendam. Disambiguating automatic semantic annotation based on a thesaurus structure. In *accepted in the 14e conference sur le Traitement Automatique des Langues Naturelles (TALN-2007)*, 2007.
10. Christian Wartena, Rogier Brussee, Luit Gazendam, and Willem-Olaf Huijsen. Apolda: A practical tool for semantic annotation. In *The 4th International Workshop on Text-based Information Retrieval (TIR 2007)*, Regensburg, Germany, September 2007.