Annotating mathematical documents: Using KAT in the context of the CorTeX Framework

Tom Wiesing

January 16, 2014

Abstract

The KAT library, short for KWARC Annotation Tool, is a tool that was and still is developed by the KWARC group at Jacobs University Bremen. It is a library for annotation documents on the web. It was originally developed for use within the CorTeX Framework. The CorTeX Framework is a framework for annotating documents.

In the introduction we will start by talking about the idea of annotating documents. While it was common practice to annotate documents manually a few years ago, this is no longer modern. The idea behind both KAT and CorTex is to digitalise and automate this. Ideally, we have a software that, given any text, annotates it completely automated, however this is not possible. There are existing systems which allow manual annotation, such as Brat, however they have the big disadvantage of only working with pure text. Another key idea behind KAT is to be able to highlight any form of data such as math or images. We will discuss in which way KAT can annotate any XML document by storing the positions of words / objects via their XPath.

Next, we will talk about the CorTex framework. As stated in the introduction, completely automatic annotating documents in not possible. The idea behind CorTex is to split up this process into several components. We first manually annotate several documents, then put them into a database and finally have a "learning" software which looks at the existing database of annotated documents and then uses that information to annotate new documents.

In the third section we will talk abou t the purpose of KAT and its place within the CorTex system. As stated before, manual annotating of documents might be neccessary to get started on a database of documents and also to correct existing annotations by other authors. This should have a nice interface for the annotators and they should not be required to install an additional program on their computer. We will also talk about the common KAT workflow. First the annotator requests a document to annotate from the system, then he / she annotates with his web browser using KAT and then sends it back to the system.

Continuing, the KAT interface and how to use KAT in its current state will be discussed. We will also discuss the idea of ontologies, which is an XML format allowing to define concepts which can be annotated with KAT. Also the specific workflow when using KAT will be discussed with the help of an example.

Furthermore, an outlook into the future development of KAT will be given. The idea of integrating KAT into the browser via a browser extension allowing annotation of any website. Also the idea of synchronising this via a centralised server will be discussed.

Finally we will wrap up the entire paper by stating what has been done already and what is still to be done.

Contents

1 Introduction

2 The CorTex framework

3 The place of KAT within CorTex

4 The KAT interface: Using KAT

5 Further development of KAT: An outlook

6   Conclusion

7 References