

KAT: an Annotation Tool for STEM Documents

Tom Wiesing, Deyan Ginev, Sourabh Lal, Michael Kohlhase

Computer Science
Jacobs University Bremen

MathUI Workshop @ CICM
Washington DC, 13. July 2015

Natural Language Processing Systems & Annotations

- ▶ parse natural language and automatically annotate documents
- ▶ commonly need a so-called “gold standard”
 - ▶ manually annotated corpus
 - ▶ used for training and evaluation
- ▶ Annotations are meta-data on top of a document
 - ▶ do not change the content of the document
 - ▶ store information about the content of the document
- ▶ need a tool to create and manage annotations

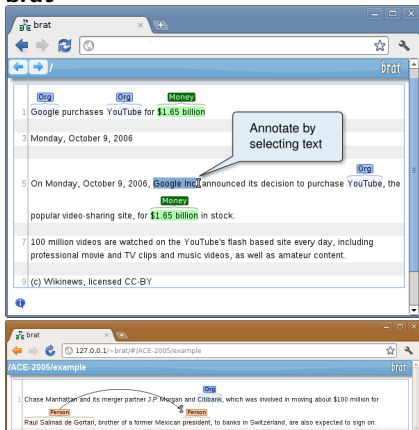
Existing Annotation Tools

- ▶ **Static Annotators**, e.g. *Annotatie*
 - ▶ annotate ranges in document “images” (e.g. scans)
- ▶ **Web Annotators**, e.g. *Hypothes.is*, *YAWAS*, ...
 - ▶ annotate online documents as well as PDFs
 - ▶ free-form comments
 - ▶ can be shared publically or privately
- ▶ **linguistic annotators**, e.g. *brat* (= brat rapid annotation tool)
 - ▶ create fix-form annotations to work with automatically
 - ▶ text based and relational annotations
 - ▶ provides pdf and html export as well as search

Observation: neither fit our purpose: linguistic annotation of STEM documents

Existing Annotation Tools: Interfaces

brat



The screenshot shows the brat web interface in a browser window. The main text area contains several lines of text with annotations. A callout box points to the text "Annotate by selecting text".

1 Google purchases YouTube for **\$1.65 billion**

3 Monday, October 9, 2006

5 On Monday, October 9, 2006, **Google Inc.** announced its decision to purchase YouTube, the popular video-sharing site, for **\$1.65 billion** in stock.

7 100 million videos are watched on the YouTube's flash based site every day, including professional movie and TV clips and music videos, as well as amateur content.

9 (c) Wikinews, licensed CC-BY

127.0.0.1/~brat/#JACE-2005/example

1 Chase Manhattan and its merger partner J.P. Morgan and Citibank, which was involved in moving about \$100 million for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

YAWAS




The screenshot shows the YAWAS web interface in a browser window. The main text area contains text from Wikipedia with annotations. A callout box points to the text "Annotate by selecting text".

Twitter is a social networking and micro-blogging service that enables its users to send and read other users' updates known as tweets. Tweets are text-based posts of up to 140 bytes in length. Updates are displayed on the user's profile page and delivered to other users who have signed up to receive them. **Subscribers can restrict delivery** to those in their circle of friends (delivery to everyone being the default). Users can send and receive updates via the Twitter website, SMS, RSS (create only), or through applications such as Tweetie, Tweetierific, Tweetbot, TweetDeck and Feedzir. The service is free to use over the web, but using SMS may incur phone services provider fees.

As of March 2009, Twitter has received extensive visibility and popularity worldwide. **Twitter is often described as the SMS of internet** in that the site provides the back-end functionality (via its APIs) to other desktop and web-based applications to send and receive short text messages often obscuring the actual website itself. This extensibility of the service has earned it more popularity than it would have gained if users had to visit the site to use the device.

Type	Private
Founded	2006
Headquarters	San Francisco, California, USA
Key people	Jack Dorsey, Chairman Evan Williams, CEO Biz Stone, Creative Director
Industry	mobile social network service, micro-blogging
Employees	347
Website	http://twitter.com/

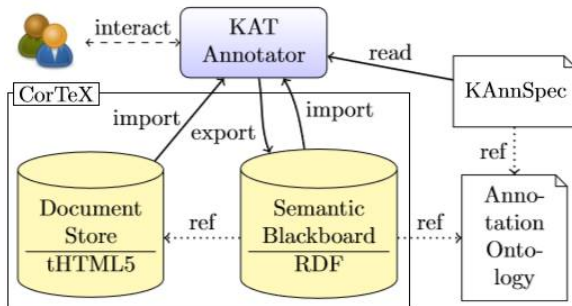
Particularities of STEM documents

- ▶ **Observation:** STEM Documents are multimodal (use more than text)
 - ▶ Formulae
 - ▶ “For each $\epsilon > 0$ ”
 - ▶ “..., then $\epsilon^2/7 > \delta...$ ”
 - ▶ “The OH group in  (ethanol)...”
 - ▶ Tables
 - ▶ Diagrams
 - ▶ Images
 - ▶ Quotes
 - ▶ Listings
 - ▶ ...
- ▶ **Observation:** XHTML5 (HTML5 in XML serialization) can represent all of them
- ▶ **KAT Approach:** annotate XHTML5 documents

What is KAT?

- ▶ KAT = **K**WARC **A**nnotation **T**ool
- ▶ web-based annotation tool (= it runs in a browser)
- ▶ stores annotations bound to ranges within XHTML5 documents (TEI-tokenized)
- ▶ has a flexible ontology specified by a **KAnnSpec**

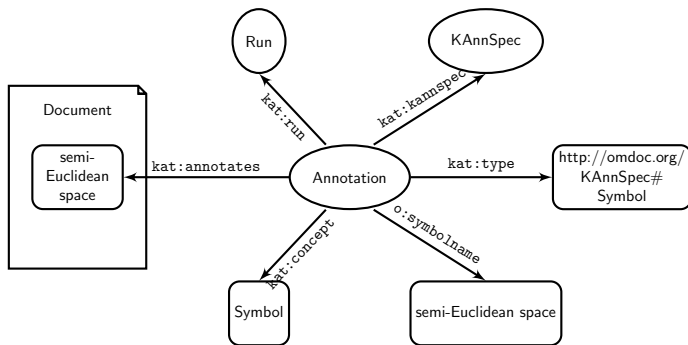
Architecture of KAT



- ▶ best used together with a Corpus Management system (e. g. CorTeX)
- ▶ annotation procedure:
 - 1) request annotation task from CorTeX
 - 2) TEI-tokenised document and KAnnSpec sent to KAT
 - 3) user annotates document
 - 4) generated annotations sent back to CorTeX

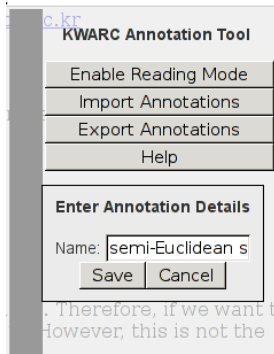
Representing Annotations in RDF

- ▶ annotation bound to an XML range in the document
 - ▶ XPath to select start and end of range
 - ▶ also point to least common ancestor element
 - ▶ gives syntax like `doc#cse(con, start, end)`
 - ▶ can only select entire XML elements (words, sentences, subformulae)
- ▶ anonymous RDF node for annotation (multiple annotations per range possible)
- ▶ RDF triples connect annotation node with annotation information



KAnnSpecs and Ontologies (1)

- ▶ KAT is not tied to a particular ontology
- ▶ we use so-called KAnnSpecs (= **KAT Annotation Specification**) that describe
 - ▶ annotation interface
 - ▶ constraints between annotation components and
 - ▶ RDF to be produced
- ▶ in particular we need to specify:
 - 1) fields of the annotation form (with values + constraints)
 - 2) their display
 - 3) RDF attributes required to create RDF triples



KAnnSpecs and Ontologies (2)

- ▶ KAnnSpec XML example
- ▶ concept declaration for an OMDoc Symbol:

```
<concept name="Symbol" rdftype="o:Symbol" >
  <documentation>An OpenMath/OMDoc Symbol</documentation>
  <field name="name" type="text" rdfpred="o:symbolname" >
    <documentation>
      The name of the symbol defines it in a theory.
    </documentation>
    <value>Name</value>
    <default>Symbol</default>
    <validation>.*</validation>
    <number atleast="1" atmost="1" />
  </field>
  <display>
    <template><b>Symbol:</b> <br/> {name}</template>
  </display>
</concept>
```

Generating RDF

- ▶ we need to be able to export the annotations
- ▶ we use RDF triples
- ▶ automatically generated from properties given in the KAnnspec
- ▶ represent the annotation and how it is linked to the document

```
<rdf:RDF xmlns:o="http://omdoc.org/KAnnspec#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:kat="https://github.com/KWARC/KAT/" >
  <!-- omitted a lot of meta-information here -->
  <rdf:Description rdf:nodeID="KAT_1433087821332_4477">
    <kat:annotates
      rdf:resource="https://kwarc.github.io/KAT/content/sample1.html#cse(%2F%2F*%5B%
%5D%2C%2F%2F*%5B%40id%3D'word.202'%5D%2C%2F%2F*%5B%40id%3D'word.203
<kat:run rdf:nodeID="kat_run" />
<kat:kannspec rdf:nodeID="KAT_1433087757661_OMDoc" />
<kat:concept>Symbol</kat:concept>
<kat:type rdf:resource="http://omdoc.org/KAnnspec#Symbol" />
<o:symbolname>semi-Euclidean space</o:symbolname>
</rdf:Description>
</rdf:RDF>
```

Time for a Demo

<https://kwarc.github.io/KAT/>

Conclusion & Future Work

- ▶ KAT is a browser-based linguistic annotation tool for STEM documents in XHTML5 form
- ▶ UI can be instantiated to a given annotation ontology via a KAnnSpec
- ▶ tested with KAnnSpecs for OMDoc, POS, Stanford grammar
- ▶ Available under GPL3 at <http://github.com/KWARC/KAT>

Conclusion & Future Work

- ▶ KAT is a browser-based linguistic annotation tool for STEM documents in XHTML5 form
- ▶ UI can be instantiated to a given annotation ontology via a KAnnSpec
- ▶ tested with KAnnSpecs for OMDoc, POS, Stanford grammar
- ▶ Available under GPL3 at <http://github.com/KWARC/KAT>
- ▶ **Future:** improve implementation
 - ▶ Better visualisation of annotations (in particular relations)
 - ▶ Optimize interactions/workflows (save clicks mouse moves)
 - ▶ Integration with CorTeX (for corpus management)
 - ▶ Test with more ontologies (KAnnSpecs) and documents
 - ▶ Stresstest in a multi-user setting

Conclusion & Future Work

- ▶ KAT is a browser-based linguistic annotation tool for STEM documents in XHTML5 form
- ▶ UI can be instantiated to a given annotation ontology via a KAnnSpec
- ▶ tested with KAnnSpecs for OMDoc, POS, Stanford grammar
- ▶ Available under GPL3 at <http://github.com/KWARC/KAT>
- ▶ **Future:** improve implementation
 - ▶ Better visualisation of annotations (in particular relations)
 - ▶ Optimize interactions/workflows (save clicks mouse moves)
 - ▶ Integration with CorTeX (for corpus management)
 - ▶ Test with more ontologies (KAnnSpecs) and documents
 - ▶ Stresstest in a multi-user setting
- ▶ **Perspective:** use KAT as a lightweight OMDoc editor (annotate and harvest)