

# AMATEXTTOOL

*Jon Holmen, Øyvind Eide, Christian-  
Emil Ore, University of Oslo, Unit for  
Digital Documentation*

*<http://www.edd.uio.no/>*

*Arve Rasmussen, PCVerkstedet ANS*

Co-funded through EPOCH, 6th framework, contract no. IST-  
2002-507382



# Outline

- ◆ Tool for semi-automatic semantic enrichment of XML tagging
- ◆ Freely available via Sourceforge
- ◆ Written in Java
- ◆ Open source: GPL
- ◆ Input format: XML
- ◆ Output format: XML



# Why

- ◆ Background: AMA production line
- ◆ For database mapping: Amatool
- ◆ Amatexttool: For semantic mapping of text documents
- ◆ A “data synthesis and modification tool”
- ◆ Based on methods used by programmers for semi-automatic tagging (e.g. Methods using Perl scripts or XSLT)



# Production line example

- ◆ Scan document, save as RTF
- ◆ Open in OpenOffice, save as TEI
- ◆ TEI document with only basic structure
- ◆ Import into Amatexttool
- ◆ Enrich tagging
- ◆ Export as XML
- ◆ Import from XML to database



# Example text as exported from OpenOffice

<p>The excavation in Wasteland in  
2005 was performed by <i>Dr.  
Diggey</i>. He had the misfortune  
of breaking the beautiful sword  
(C50435) in 30 pieces.</p>



# How - method 1

- ◆ Search for patterns
- ◆ Find mapping equivalent entities in CIDOC-CRM (eg. person name --> E82 Actor appellation, place name --> E44 Place appellation)
- ◆ Insert CIDOC-CRM compliant XML markup as batch operations



# Example text after tagging

<p>The <event>excavation in  
<placeName>Wasteland</placeName>  
in <date>2005</date></event> was  
performed by <i><persName>Dr.  
Diggey</persName></i>. <rs>He</rs>  
had the misfortune of breaking the  
beautiful sword (C50435) in 30  
pieces.</p>



# How - method 2

- ◆ Identify co-reference within document (eg. this person reference represents the same real world person as that person reference, as “Dr. Diggey” and “He”)
- ◆ Insert XML id-idref’s identifying such related entities





# How - method 3

- ◆ Find mapping equivalent properties (eg. relations between persons and places, such as “Dr. Diggey” and “Wastland” connected through events)
- ◆ Insert CIDOC-CRM compliant XML markup as batch operations
- ◆ Export XML document



# Content management

- ◆ XML DTD or schema based encoding
- ◆ Include CIDOC-CRM compliant tagset in TEI through the ODD system

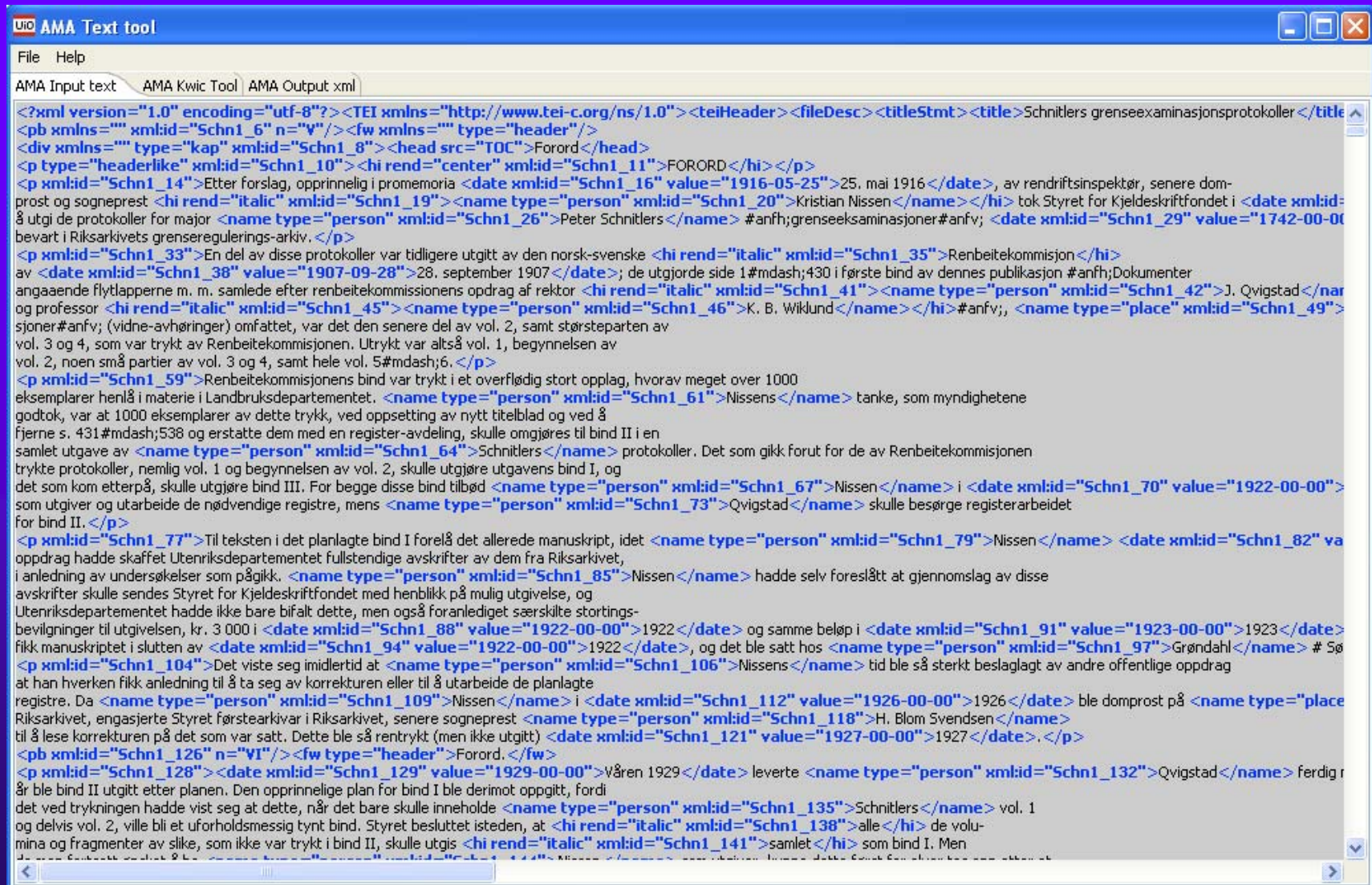


# Search system

- ◆ Free text searches
- ◆ Regular expressions
- ◆ Interface for user-specified search procedures to be included as java classes



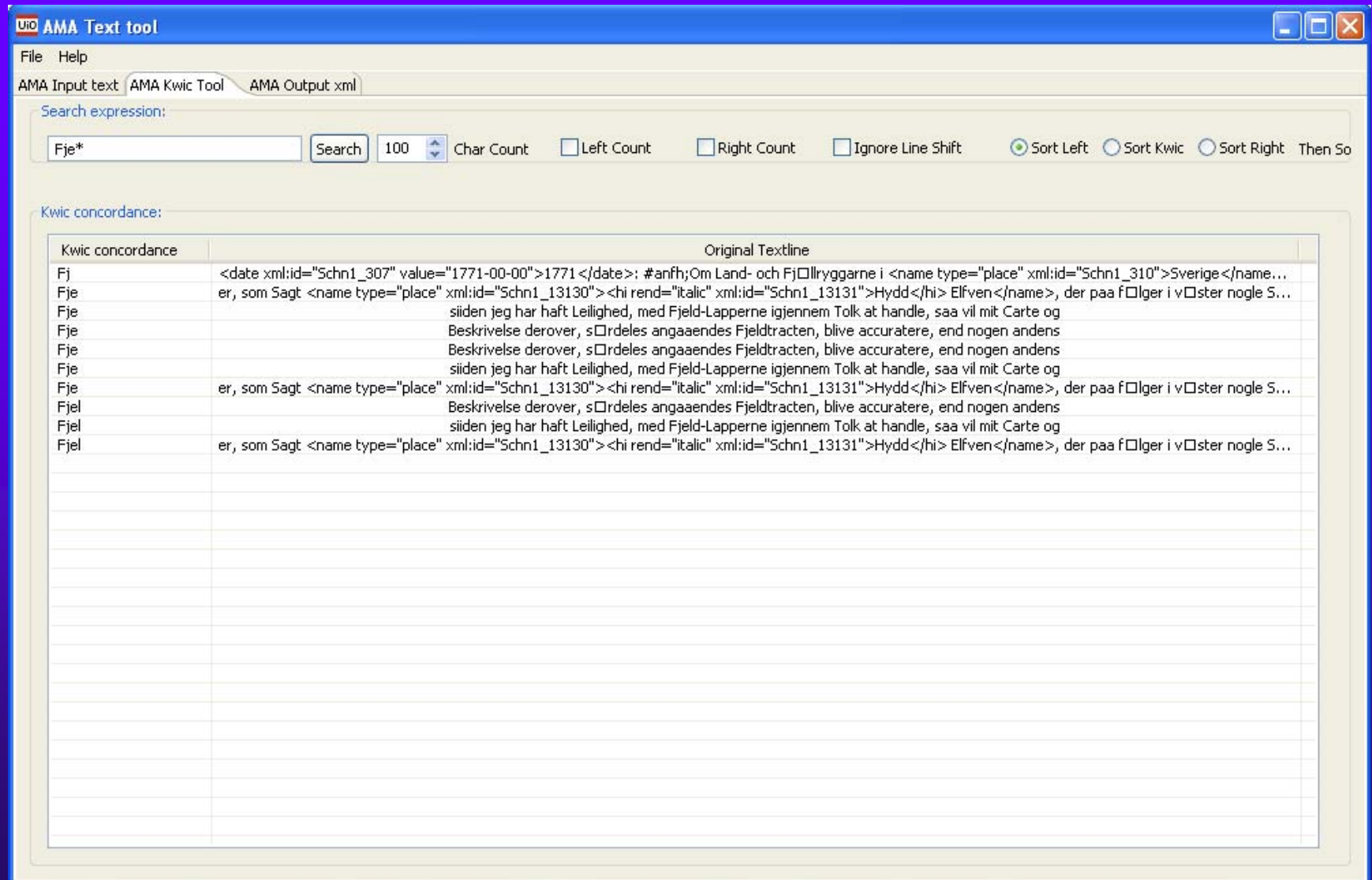
# XML input view



```
<?xml version="1.0" encoding="utf-8"?><TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader><fileDesc><titleStm><title>Schnitlers grenseexaminasjonsprotokoller</title>
<pb xmlns="" xmlid="Schn1_6" n="V"/><fw xmlns="" type="header"/>
<div xmlns="" type="kap" xmlid="Schn1_8"><head src="TOC">Forord</head>
<p type="headerlike" xmlid="Schn1_10"><hi rend="center" xmlid="Schn1_11">FORORD</hi></p>
<p xmlid="Schn1_14">Etter forslag, opprinnelig i promemoria <date xmlid="Schn1_16" value="1916-05-25">25. mai 1916</date>, av rendriftsinspektør, senere dom-
prost og sogneprest <hi rend="italic" xmlid="Schn1_19"><name type="person" xmlid="Schn1_20">Kristian Nissen</name></hi> tok Styret for Kjeldeskriftfondet i <date xmlid="
& utgi de protokoller for major <name type="person" xmlid="Schn1_26">Peter Schnitlers</name> #anfhy;grenseeksaminasjoner#anfvy; <date xmlid="Schn1_29" value="1742-00-00">
bevart i Riksarkivets grenseregulerings-arkiv.</p>
<p xmlid="Schn1_33">En del av disse protokoller var tidligere utgitt av den norsk-svenske <hi rend="italic" xmlid="Schn1_35">Renbeitekommisjon</hi>
av <date xmlid="Schn1_38" value="1907-09-28">28. september 1907</date>; de utgjorde side 1#mdash;430 i første bind av dennes publikasjon #anfhy;Dokumenter
angaaende flytlapperne m. m. samlede efter renbeitekommissionens opdrag af rektor <hi rend="italic" xmlid="Schn1_41"><name type="person" xmlid="Schn1_42">J. Qvigstad</nar
og professor <hi rend="italic" xmlid="Schn1_45"><name type="person" xmlid="Schn1_46">K. B. Wiklund</name></hi>#anfvy;, <name type="place" xmlid="Schn1_49">
sjoner#anfvy; (vidne-avhøringer) omfattet, var det den senere del av vol. 2, samt størsteparten av
vol. 3 og 4, som var trykt av Renbeitekommissionen. Uttrykt var altså vol. 1, begynnelsen av
vol. 2, noen små partier av vol. 3 og 4, samt hele vol. 5#mdash;6.</p>
<p xmlid="Schn1_59">Renbeitekommissionens bind var trykt i et overflødig stort opplag, hvorav meget over 1000
eksemplarer henlå i materie i Landbruksdepartementet. <name type="person" xmlid="Schn1_61">Nissens</name> tanke, som myndighetene
godtok, var at 1000 eksemplarer av dette trykk, ved oppsetning av nytt titelblad og ved å
fjerne s. 431#mdash;538 og erstatte dem med en register-avdeling, skulle omgjøres til bind II i en
samlet utgave av <name type="person" xmlid="Schn1_64">Schnitlers</name> protokoller. Det som gikk forut for de av Renbeitekommissionen
trykte protokoller, nemlig vol. 1 og begynnelsen av vol. 2, skulle utgjøre utgavens bind I, og
det som kom etterpå, skulle utgjøre bind III. For begge disse bind tilbød <name type="person" xmlid="Schn1_67">Nissen</name> i <date xmlid="Schn1_70" value="1922-00-00">
som utgiver og utarbeide de nødvendige registre, mens <name type="person" xmlid="Schn1_73">Qvigstad</name> skulle besørge registerarbeidet
for bind II.</p>
<p xmlid="Schn1_77">Til teksten i det planlagte bind I forelå det allerede manuskript, idet <name type="person" xmlid="Schn1_79">Nissen</name> <date xmlid="Schn1_82" va
opdrag hadde skaffet Utenriksdepartementet fullstendige avskrifter av dem fra Riksarkivet,
i anledning av undersøkelser som pågikk. <name type="person" xmlid="Schn1_85">Nissen</name> hadde selv foreslått at gjennomslag av disse
avskrifter skulle sendes Styret for Kjeldeskriftfondet med henblikk på mulig utgivelse, og
Utenriksdepartementet hadde ikke bare bifalt dette, men også foranlediget særskilte stortings-
bevilgninger til utgivelsen, kr. 3 000 i <date xmlid="Schn1_88" value="1922-00-00">1922</date> og samme beløp i <date xmlid="Schn1_91" value="1923-00-00">1923</date>
fikk manuskriptet i slutten av <date xmlid="Schn1_94" value="1922-00-00">1922</date>, og det ble satt hos <name type="person" xmlid="Schn1_97">Grøndahl</name> # Sø
<p xmlid="Schn1_104">Det viste seg imidlertid at <name type="person" xmlid="Schn1_106">Nissens</name> tid ble så sterkt beslaglagt av andre offentlige oppdrag
at han hverken fikk anledning til å ta seg av korrekturen eller til å utarbeide de planlagte
registre. Da <name type="person" xmlid="Schn1_109">Nissen</name> i <date xmlid="Schn1_112" value="1926-00-00">1926</date> ble domprost på <name type="place"
Riksarkivet, engasjerte Styret førstearkivar i Riksarkivet, senere sogneprest <name type="person" xmlid="Schn1_118">H. Blom Svendsen</name>
til å lese korrekturen på det som var satt. Dette ble så rentrykt (men ikke utgitt) <date xmlid="Schn1_121" value="1927-00-00">1927</date>.</p>
<pb xmlid="Schn1_126" n="VI"/><fw type="header">Forord.</fw>
<p xmlid="Schn1_128"><date xmlid="Schn1_129" value="1929-00-00">Våren 1929</date> leverte <name type="person" xmlid="Schn1_132">Qvigstad</name> ferdig r
år ble bind II utgitt etter planen. Den opprinnelige plan for bind I ble derimot oppgitt, fordi
det ved trykningen hadde vist seg at dette, når det bare skulle inneholde <name type="person" xmlid="Schn1_135">Schnitlers</name> vol. 1
og delvis vol. 2, ville bli et uforholdsmessig tynt bind. Styret besluttet isteden, at <hi rend="italic" xmlid="Schn1_138">alle</hi> de volu-
mina og fragmenter av slike, som ikke var trykt i bind II, skulle utgis <hi rend="italic" xmlid="Schn1_141">samlet</hi> som bind I. Men
dette forutsettes ved at <name type="person" xmlid="Schn1_144">Nissen</name> og <name type="person" xmlid="Schn1_145">Qvigstad</name> skulle sørge for at de nødvendige registre
for bind I ble utarbeidet og trykt sammen med boken.
```



# Search and select



The screenshot shows the 'AMA Text tool' window. The search expression is 'Fje\*'. The search results are displayed in a table with two columns: 'Kwic concordance' and 'Original Textline'.

Kwic concordance	Original Textline
Fj	<date xml:id="Schn1_307" value="1771-00-00">1771</date>: #anhf;Om Land- och Fj�llyggarne i <name type="place" xml:id="Schn1_310">Sverige</name...
Fje	er, som Sagt <name type="place" xml:id="Schn1_13130"><hi rend="italic" xml:id="Schn1_13131">Hydd</hi> Elfven</name>, der paa f�lger i v�ster nogle S...
Fje	siiden jeg har haft Leilighed, med Fjeld-Lapperne igjennem Tolk at handle, saa vil mit Carte og
Fje	Beskrivelse derover, s�rdeles angaaendes Fjeldtracten, blive accuratere, end nogen andens
Fje	Beskrivelse derover, s�rdeles angaaendes Fjeldtracten, blive accuratere, end nogen andens
Fje	siiden jeg har haft Leilighed, med Fjeld-Lapperne igjennem Tolk at handle, saa vil mit Carte og
Fje	er, som Sagt <name type="place" xml:id="Schn1_13130"><hi rend="italic" xml:id="Schn1_13131">Hydd</hi> Elfven</name>, der paa f�lger i v�ster nogle S...
Fjel	Beskrivelse derover, s�rdeles angaaendes Fjeldtracten, blive accuratere, end nogen andens
Fjel	siiden jeg har haft Leilighed, med Fjeld-Lapperne igjennem Tolk at handle, saa vil mit Carte og
Fjel	er, som Sagt <name type="place" xml:id="Schn1_13130"><hi rend="italic" xml:id="Schn1_13131">Hydd</hi> Elfven</name>, der paa f�lger i v�ster nogle S...





# Conclusions

- ◆ Tool will be ready for testing by February 2008
- ◆ We will use it internally for encoding of e.g. lexicographical material
- ◆ Will be distributed to EPOCH partners and beyond
- ◆ Email to [oyvind.eide@edd.uio.no](mailto:oyvind.eide@edd.uio.no) to be put on the mailing list

