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

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

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

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

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

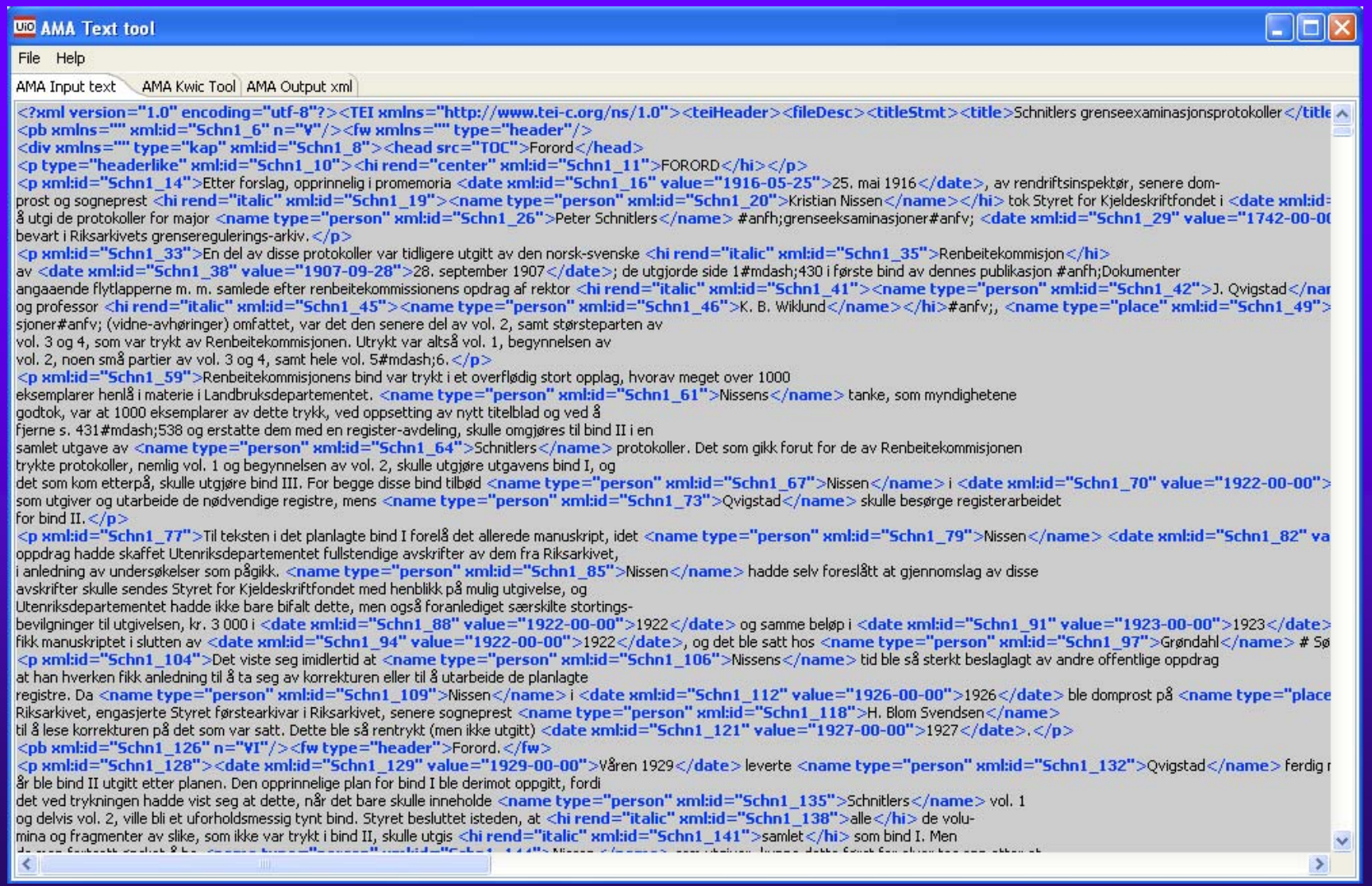
Content management

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

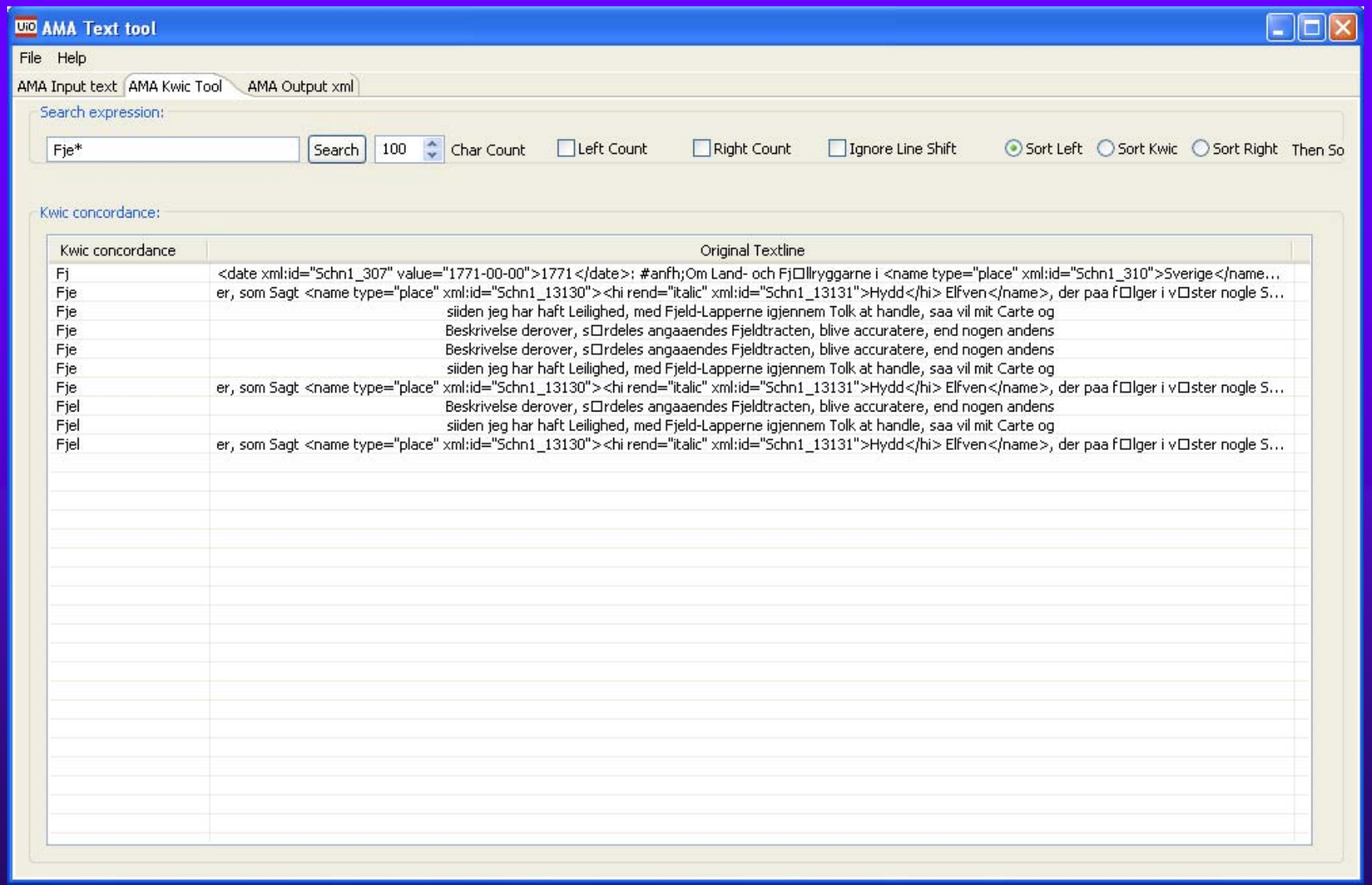
Search system

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

XML input view



Search and select



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 to be put on the mailing list