

(Some) Standards in the Humanities

Sebastian Drude
CLARIN ERIC

RDA 4th Plenary, Amsterdam
September 2014

Overview

1. Introduction
2. Written text: the Text Encoding Initiative (TEI)
3. Multimodal: ELAN Annotation Format (EAF)
4. Lexical Data: Lexical Markup Framework (LMF)
5. Metadata: Component Metadata Infrastructure (CMDI)
6. Conclusions

1) Introduction

- Many formats, almost no real 'standards' (ISO)
- Different kinds of standards:
 1. One clear limited purpose, one suitable format
 2. Integration of many purposes and formats
 - A. One all-inclusive extensive and extendable standard
 - B. A 'container' format with heterogeneous content
 - C. Playing 'Lego': flexible combining of building blocks
- General challenge: Semantic interoperability

Overview

1. Introduction
2. **Written text: the Text Encoding Initiative (TEI)**
3. Multimodal: ELAN Annotation Format (EAF)
4. Lexical Data: Lexical Markup Framework (LMF)
5. Metadata: Component Metadata Infrastructure (CMDI)
6. Conclusions

2) Text Encoding Initiative (TEI)

- Mark-up of written text
- Goals, for example:
 - Critical editions
 - making the structure explicit
 - labelling of features/elements of interest
- TEI – an international consortium since 1987
- <http://www.tei-c.org/>
- Set of Guidelines (now 'P5', released 2007)

2) Text Encoding Initiative (TEI)

- P5: Now more than 1600 pages
- TEI-Guidelines usually implemented as XML
- Available as DTD/ODD, also in <oXygen/>
- Individual chapters for special text types, parts and entities (poetry, persons, writing systems)
- Applications use TEI-compliant subsets
- But also non-standard extensions are used

2) Text Encoding Initiative (TEI)



University of Oxford Text Archive

University of Oxford Text Archive: [Home](#) | [About](#) | [Catalogue](#) | [Contact](#) | [Help and FAQ](#) | [News](#) | [Search OTA](#)

The Auchinleck Manuscript

Title	The Auchinleck Manuscript [Electronic resource]
Author	Burnley, David ; Wiggins, Alison ; National Library of Scotland
Availability	Distributed by the University of Oxford under a Creative Commons Attribution-NonCommercial
	Download: zip
Languages	English, Middle (1100-1500)
Editorial Practice	Encoding format: TEI XML
OTA keywords	Romances
LC keywords	Literature, Medieval Romances

```

<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Alphabetical Praise of Women</title>
        <author>Auchinleck Project</author>
      </titleStmt>
      <publicationStmt>
        <p>This file is part of the facsimile-edition of the Auchinleck Manuscript was
          co-edited by Professor David Burnley and Dr Alison Wiggins. The HTML ve
          the resou
        </publicationStmt>
      <notesStmt>
        <note>This manuscr
          conver
          notes
          a sepa
        </note>
        <p>
          <hi rend="bo
        </p>
        </note>
        <note>
          <p>11-line stan
            <emph>Sir Tristrem</emph>; rhy
            some defective through cutting
            lost the first 17 lines.</p>
          </note>
          <note>
            <p>Unique copy.
          </note>
        </notesStmt>
      <sourceDesc>
        <p>The Auchinleck Manuscript (NLS Adv MS 19.2.1) is one of th
          Scotland's greatest treasures. Produced in London
          unique insight into the English language and literat
          generation grew up with and were influenced by. It a
          first known owner, Lord Auchinleck, who discovered t
          donated it to the precursor of the National Library
        </p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <head>Alphabetical Praise of Women</head>
      <l>.... <milestone unit="folio" n="f.324ra"/>
        <note type="xref">
          <ref target="#324r.sid"/>
        </note>
      </l>
      <l>.... </l>
      <l>.... </l>
      <l>.... </l>
      <l n="5">....
      .... </l>
      <l>Bot fals men make her fingres feld</l>
      <l>& do<c type="thorn">#x00FE;</c> hem wepe wel sore, to re
      <l>Her res:</l>
      <l n="10"><c type="THORN">#x00FE.</c>urth

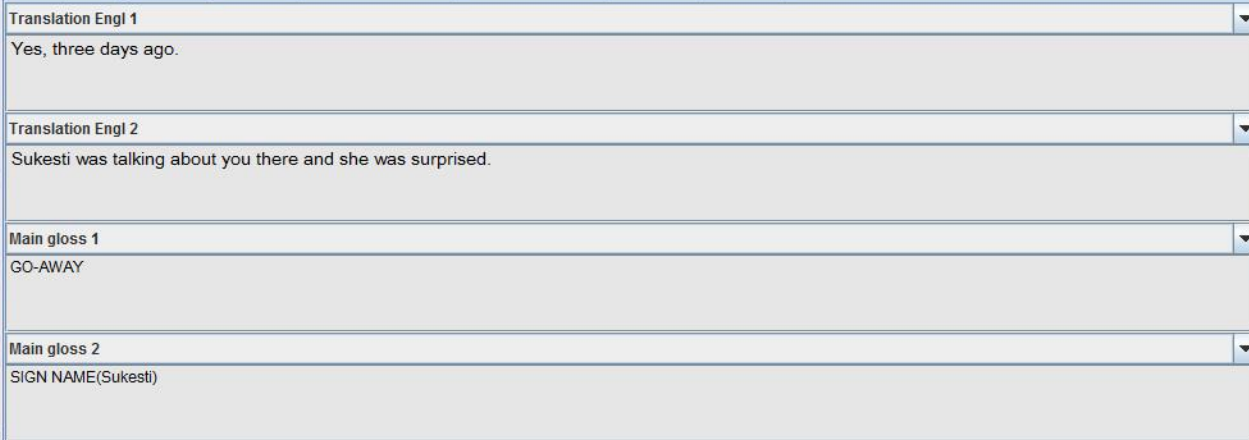
```


Overview

1. Introduction
2. Written text: the Text Encoding Initiative (TEI)
3. **Multimodal: ELAN Annotation Format (EAF)**
4. Lexical Data: Lexical Markup Framework (LMF)
5. Metadata: Component Metadata Infrastructure (CMDI)
6. Conclusions

3) ELAN Annotation Format (EAF)

- Annotation of audio and video recordings
- Spoken language, multimodal communication
- Fundamental: time-relatedness
- Stand-off annotation to the original recordings
- ELAN tool for linguistic annotation by [MPI-PL](#)
- www.mpi.nl/tools/elan/EAF_Annotation_Format.pdf
- Horizontal: Time stamps and segments
- Vertical: tiers for different information types and speakers



Selection: 00:00:08.523 - 00:00:32.043 23520



Example 1	
2014-09-23, Amsterdam	S. Drude @ RDA Plenary 4
	11

Overview

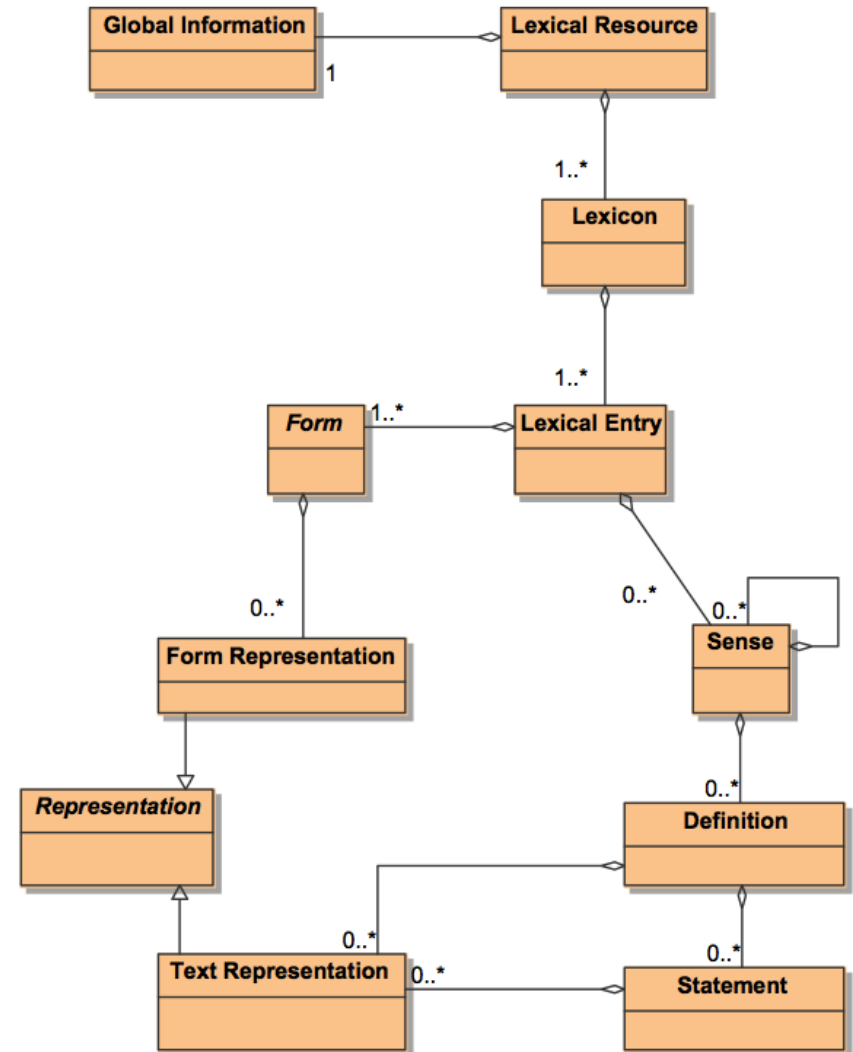
1. Introduction
2. Written text: the Text Encoding Initiative (TEI)
3. Multimodal: ELAN Annotation Format (EAF)
4. **Lexical Data: Lexical Markup Framework (LMF)**
5. Metadata: Component Metadata Infrastructure (CMDI)
6. Conclusions

4) Lexical Markup Framework (LMF)

- Many kinds of “dictionary”-type data:
 - linguistic information on words/forms/expressions
 - encyclopaedic information on things/persons/...
- Very many different structures, granularity, ...
- Need for aggregation of different sources
- ISO TC 37/SC 4/WG 4: [ISO 24613:2008](#)
- www.lexicalmarkupframework.org

4) Lexical Markup Framework (LMF)

- Core UML model (→)
- Only this is obligatory
- Feature structures (TEI)
- Connection with [ISOcat](#)
- Several extensions are recommended but not part of the standard
- In principle, anything can be added
- Weak: “LMF-compliant”??



Overview

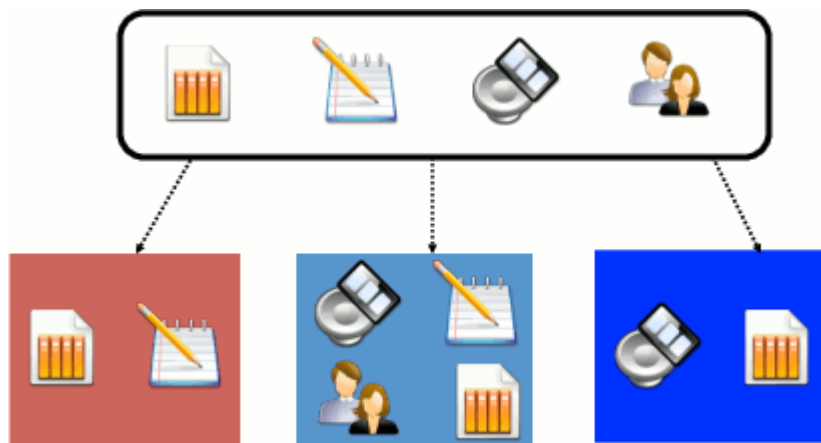
1. Introduction
2. Written text: the Text Encoding Initiative (TEI)
3. Multimodal: ELAN Annotation Format (EAF)
4. Lexical Data: Lexical Markup Framework (LMF)
5. **Metadata: Component Metadata Infrastructure (CMDI)**
6. Conclusions

5) Component Metadata Infrastructure (CMDI)

- Metadata are key for any infrastructure
- Very many different needs, heterogeneous formats and elements / metadata categories
- A flexible standard is needed
- CMDI: developed within **CLARIN** – the Common Language Resources & technology INfrastructure
- www.clarin.eu/content/component-metadata
- Introduction & overview:
www.clarin.nl/node/146

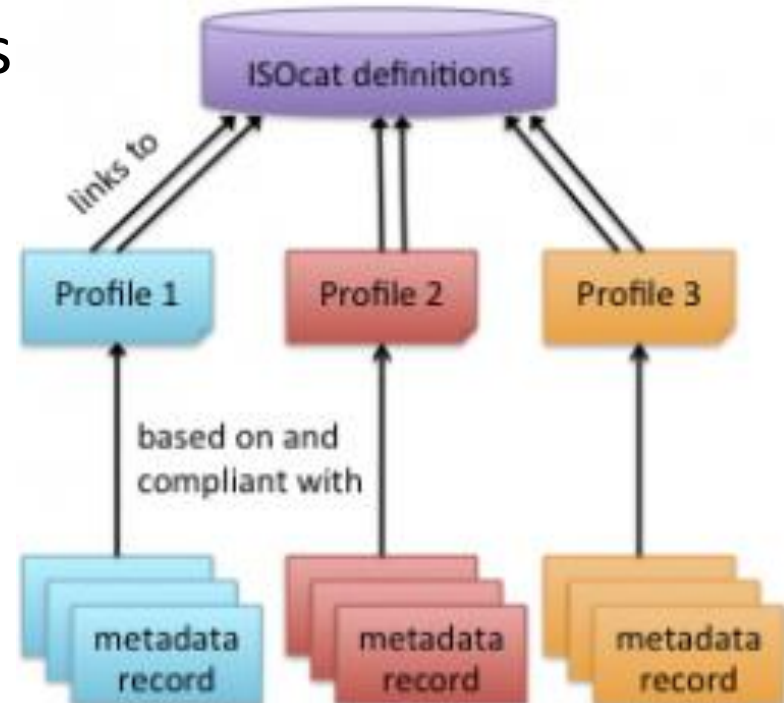
5) Component Metadata Infrastructure (CMDI)

- CMDI: components / blocks of elements that can be enhanced and combined to *profiles*
- Components & profiles are registered for re-use
- All MD elements and concepts are linked to [ISOcat](#)



Components

Profiles



Overview

1. Introduction
2. Written text: the Text Encoding Initiative (TEI)
3. Multimodal: ELAN Annotation Format (EAF)
4. Lexical Data: Lexical Markup Framework (LMF)
5. Metadata: Component Metadata Infrastructure (CMDI)
6. **Conclusions**

6) Conclusions

- Different purposes call for different solutions
- TEI: huge comprehensive schema, selections
- EAF: limited focused stable stand-off markup
- LMF: meta-container, heterogeneous content
- CMDI: components combined in profiles
- All are currently implemented in XML
- All (can) refer to [ISOcat](#) (TEI: feature structures)
- TEI & EAF are dominant, LMF is least used

(Some) Standards in the Humanities

Sebastian Drude
CLARIN ERIC

RDA 4th Plenary, Amsterdam
September 2014