XML for quantitative codicology

Matthew James Driscoll
Silvia Hufnagel
Eric Andrew Haswell

The Arnamagnæan Institute
University of Copenhagen

27 January 2012
MRSH, Université de Caen Basse-Normandie
Séminaire: Exploration et exploitation des sources encodées en TEI
Background

Stories for all time: The Icelandic fornaldarsögur
From March 2011, the four-year research project, based at the Arnamagnæan Institute, University of Copenhagen, will investigate the transmission of one of the more popular saga genres from the middle ages onwards, mapping their production, dissemination and reception in relation to broader historical, social and cultural processes.

36 sagas, of which there are 1542 individual texts, preserved in 779 manuscripts.


Two two-year post-doc positions from March 2012.
Theoretical background: Quantitative codicology


- Explored the research possibilities afforded by collecting a large amount of statistical data on the physical characteristics of medieval codices.


Bibliography and the sociology of texts

D. F. McKenzie’s 1985 Panizzi lectures, published as *Bibliography and the sociology of texts* (London, 1986), argued that any history of the book ‘must take into account the social, economic and political motivations of publishing, the reasons why texts were written and read as they were, why they were rewritten and redesigned, or allowed to die’.

Bibliography, McKenzie said, is ‘the study of the sociology of texts’ because it deals with ‘the human motives and interactions which texts involve at every stage of their production, transmission and consumption’.
‘New’ or ‘material’ philology


The so-called ‘new’ or ‘material’ philology, with inspiration from Bernard Cerquiglini’s polemical essay *Éloge de la variante : Histoire critique de la philologie* (Paris, 1989), essentially argued that texts cannot be dissociated from their material embodiments—that the physical form of a text is part of its meaning—or from the cultural, historical and ideological forces at work in their production, dissemination and reception.
The ‘artefactual’ turn

We seek to achieve a better understanding of the structure and mechanisms of the production, dissemination and reception of chirographically transmitted texts, and in doing so, to develop a philological praxis which views texts not as abstracts but as artefacts, material objects shaped and reshaped through human agency.

We wish to describe and transcribe these artefacts as carefully as possible, and also to link them to other artefacts preserving texts of the same (and other) works.

More importantly, we wish to map the relationships between these artefacts and the people who produced and consumed them, to show in a dynamic way how the ‘manuscript matrix’ worked.
XML schema

Very restrictive subset of TEI P5 XML for manuscript description.

Schema is specified using the TEI's ODD system and ROMA suite of tools which allow the schema and accompanying documentation to be simultaneously produced.

Schema designed with a view towards eliminating natural language as much as possible from XML.

Manuscript features distilled down to a set of core features, which are expected to be present in every manuscript description.

This approach to schema design results in a set of data about different manuscripts which can be easily compared and analysed.
Encoding example: `<supportDesc>`

The content model of `supportDesc` is illustrative of some aspects of schema customisation:

```xml
<supportDesc material="chart">
  <support>
    <num type="front-flyleaf" value="3"/>
    <num type="book-block" value="201"/>
    <num type="back-flyleaf" value="3"/>
    <dimensions type="leaf">
      <height max="195" min="205" unit="mm"/>
      <width quantity="150" unit="mm"/>
    </dimensions>
  </support>
  <foliation ana="#FOL_contemporary"/>
  <condition ana="#CON_average"/>
</supportDesc>
```
Examples of encoded characteristics

In addition to encoding aspects of a manuscript's support, the schema also mandates that many other data characteristics be supplied in a similarly data-intensive fashion. These include:

- the number of columns
- the number of written lines
- the number of words per line
- dimensions of the written area
- number of hands in the manuscript and relative scope of each
- the names of the scribes identified as corresponding to hands in the manuscript
- the relative level of decoration of the manuscript
- the relative level of additions (e.g. marginal comments) made to the manuscript
- the degree to which the binding is decorated and the contemporaneousness of the binding with the manuscript
- the date and place of origin
- the names of previous owners or other individuals known to have had a part in the manuscript's history
Table of contents

1. Sörla saga sterka and its manuscripts
2. Quantitative codicology:
   written area – page size
   abbreviations
   text density
3. Scholarly manuscripts
4. Non-scholarly manuscripts
5. Conclusions
**Sörla saga sterka and its manuscripts**

*Fornaldarsaga norðurlanda* (ancient saga of the Northern lands)

Sörli the Strong fights against black men, giants and King Hálfdan Brönumfóstri and his family, ends in triple wedding

3 versions: A (short), B (long), C (very long)

2 editions:
- Biörner *Nordiska kämpa dater* (Sth, 1737)
- Rafn *Fornaldar sögur nordrlanda* (Cph, 1829-30)

8 sets of *rímur*
Sørla saga sterka and its manuscripts

31 manuscripts ss. XVII-XX in

Scholarly manuscripts (blue)
written for or by scholars

Non-scholarly manuscripts
  tax value of farms, biographical information
  very wealthy scribes (green)
  impecunious scribes (red)

symbols of status
prestigious reading mss
“paperback manuscripts”
Quantitative codicology

Based on TEI P5-conformant XML-transcriptions of all texts

Written area – page size

Number of abbreviation per metre (1m)

Text density: number of characters per square metre (1m²)
Scholarly manuscripts: written area – page size
Scholarly manuscripts: number of abbreviations
Scholarly manuscripts: text density
Non-scholarly manuscripts: symbols of status

BL Add 4857
(1669-70, for Magnús Jónsson í Vigur)

Large margins
Rather high no of abbreviations
Rather low text density
More decoration (title page, half-diamond indentation, decorated initials)

Material and textual aspects: manuscript emphasises wealth of commissioner
Non-scholarly manuscripts: prestigious reading mss

JS 11 8vo (1780, perhaps by Þorkell Jónsson)

Small margins
Large no of abbreviations
High text density
More decoration (colours, decorated initials)

Material aspects affirm conclusion of textual analysis: function of entertainment
Non-scholarly manuscripts: “paperback manuscripts”

ÍB 277 4to (1833-34, Gunnlaugur Jónsson á Skuggabjörgum)

Average margins
Average no of abbreviations
Average text density
No decoration

Material aspects affirm conclusion of textual analysis: text more important than material
Conclusions

Scholarly manuscripts:
- Large margins
- Few abbreviations
- Low text density

Reflect needs of scholars, work tools

Non-scholarly manuscripts:
- Symbols of status, heavily decorated
- Prestigious reading manuscripts
- “Paperback manuscripts”, plain reading mss

Reflect socio-economic status of scribe or commissioner
XML in a Database

Distinction between data-centric and document-centric data structures

- Many manuscript descriptions are a mix of these.

Relational databases (e.g. MySQL, PostgreSQL) are highly unsuited for document-centric data

We employ the native XML database eXist

- XML document is the database record
- No XML modifications are necessary prior to ingestion

Query language for XML: XQuery

- W3C recommendation
- Similar capabilities to SQL
Database/Server structure
Processing and analysis

• Using XQuery, we are able to query a collection of XML documents in eXist and return data in a format suitable for analysis.
• This format may be one of several options depending on how we are undertaking our analysis.
The SIMILE framework: data extraction

• Using the SIMILE widget Exhibit, we are able to plot data by space and time, and to provide a faceted browser for data exploration.

• SIMILE expects data in JSON format, and we employ XQuery to produce this data.

```{ items: [  
{  
  label: "Jón Erlendsson (JonErl001)",  
  relName: "scribe",  
  relFrom: "1639",  
  relTo: "1672",  
  relCentury: "1600",  
  personID: "JonErl001",  
  personName: "Jón Erlendsson",  
  birth: ",",  
  death: "1672",  
  socec: "Socio-economic status 1",  
  mss: [ "AM02-0009" ],  
  placeID: "VilVil01",  
  settlement: "Villingaholt",  
  settlementType: "farm",  
  addressLatLng: "63.883997,-20.750909"  
} ] };
```
The SIMILE framework: visualisation
Socio-economic status of scribes

<table>
<thead>
<tr>
<th>Status</th>
<th>17th</th>
<th>18th</th>
<th>19th</th>
<th>20th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>17%</td>
<td>0%</td>
<td>13%</td>
<td>0%</td>
</tr>
<tr>
<td>Y</td>
<td>17%</td>
<td>50%</td>
<td>38%</td>
<td>0%</td>
</tr>
<tr>
<td>XB</td>
<td>8%</td>
<td>6%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>VI</td>
<td>4%</td>
<td>13%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>V</td>
<td>8%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>IV</td>
<td>21%</td>
<td>6%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>III</td>
<td>4%</td>
<td>13%</td>
<td>13%</td>
<td>100%</td>
</tr>
<tr>
<td>II</td>
<td>0%</td>
<td>0%</td>
<td>13%</td>
<td>0%</td>
</tr>
<tr>
<td>I</td>
<td>21%</td>
<td>13%</td>
<td>25%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Economy and material

Text density and commissioner status

<table>
<thead>
<tr>
<th>Year</th>
<th>Text Density</th>
<th>Commissioners</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>20,70497347</td>
<td>5,142857143</td>
</tr>
<tr>
<td>1700</td>
<td>19,37867778</td>
<td>4</td>
</tr>
<tr>
<td>1800</td>
<td>18,22063256</td>
<td>2</td>
</tr>
<tr>
<td>1900</td>
<td>17,5148729</td>
<td>3</td>
</tr>
</tbody>
</table>
Other possibilities

1. Incorporating data from other sources
   - Data which is collected according to the same principles may be easily integrated, regardless of its source format

2. Merging data-centric and document centric manuscript data into a single file
   - Introduce a less restrictive schema designed for detailed descriptions and/or transcriptions.
   - Integrate all XML-encoded information in a single file
Thanks for your attention

Matthew James Driscoll
mjd@hum.ku.dk

Silvia Hufnagel
silvia@hum.ku.dk

Eric Andrew Haswell
eah@hum.ku.dk