



# SGML / XML & Markup

Stephen Miller  
AAC / ÖAW  
Texts & Files WS 2005–06  
Universität Wien

# DWDS

- Das Ergebnis der Corpuserstellung sind zwei verschiedene Corpora:
- Das **Kerncorpus**: Größe: 100 Millionen Textwörter
  - ausgewogene Textgrundlage
  - rechtlich abgesicherte Nutzungsvereinbarungen
  - annotiert gemäß XML/TEI
- Das **Ergänzungscorpus**: 980 Millionen Textwörter
  - opportunistisch

[http://www.dwds.de/pages/pages\\_info/dwds\\_info.htm](http://www.dwds.de/pages/pages_info/dwds_info.htm)

2 Texts & Files WS 2005–06 / SGML XML & Markup © AAC / ÖAW 2005

# British National Corpus

- The corpus is encoded according to the *Guidelines of the [Text Encoding Initiative](#) (TEI)*, using ISO standard 8879 ( [SGML: Standard Generalized Markup Language](#)) to represent both the output from CLAWS and a variety of other structural properties of texts (e.g. headings, paragraphs, lists etc.). Full classification, contextual and bibliographic information is also included with each text in the form of a TEI-conformant header.

3 Texts & Files WS 2005–06 / SGML XML & Markup © AAC / ÖAW 2005

# American National Corpus

- Encoding of the ANC data**
  - The [ANC](#) will be encoded according to the XML specifications of the [XCES](#), which specifies a flexible document structure suitable for delivery on the World Wide Web and allows for "layered" annotation documents that can be added incrementally at later stages. In particular, the data will be provided in a "stand-off" format, where annotations are provided in XML documents separate from and linked to the primary data.

4 Texts & Files WS 2005–06 / SGML XML & Markup © AAC / ÖAW 2005

# Steven DeRose

- Coombs, James H., Allen H. Renear, and Steven J. DeRose. 1987. "Markup Systems and the Future of Scholarly Text Processing." *Communications of the ACM* 30 (11): 933-947.
- DeRose Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3-26.

5 Texts & Files WS 2005–06 / SGML XML & Markup © AAC / ÖAW 2005



# What is Markup?

Stephen Miller  
AAC / ÖAW  
Texts & Files WS 2005–06  
Universität Wien

## Concepts of Markup

- Presentational
- Procedural
- Descriptive
  - James H. Coombs, Allen H. Renear, Steven De Rose, "Markup Systems and the Future of Scholarly Text Processing," *Communications of the ACM*, 30 (1987): 933-47.
  - <http://www.oasis-open.org/cover/coombs.html>

7 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005

## Types of Markup

1. Presentational
2. Procedural
3. Descriptive

8 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005

## SEASONS AND MONTHS

*The New Year*

I. Manx, Yn Vlein Noa. *The New Year*. January 1st. "... anciently the first day of November was the first day of the year: and the Mummers on the eve of All Saints' Day still begin their petition with these remarkable words, 'To-night is New Year's night. Og-u-naa. The Moon shines fair and bright. Tro-la-la.'" Kelly (*circa* 1790), s.v. Baal-sauin. See November 1st.

9 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005

## Typescript

Seasons and Months

The New Year

I. Manx, Yn Vlein Noa. *The New Year*. January 1st. "... anciently the first day of November was the first day of the year: and the Mummers on the eve of All Saints' Day still begin their petition with these remarkable words, 'To-night is New Year's night. Og-u-naa. The Moon shines fair and bright. Tro-la-la.'" Kelly (*circa* 1790), s.v. Baal-sauin. See November 1st.

10 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005

## Presentational

Season

Header1: uppercase

Header2: italic / centred

Use Adobe Garamond + Old Style figures

I. Manx, Yn Vlein Noa. *The New Year*. January 1st. "... anciently the first day of November was the first day of the year: and the Mummers on the eve of All Saints' Day still begin their petition with these remarkable words, 'To-night is New Year's night. Og-u-naa. The Moon shines fair and bright. Tro-la-la.'" Kelly (*circa* 1790), s.v. Baal-sauin. See November 1st.

Body text: 10/12

Header1 12/14

Header2 10/12

Footnotes 9/10

11 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005

## Procedural

```



12 Texts & Files WS 2005-06 / SGML XML & Markup © AAC / ÖAW 2005


```



## Descriptive

```
<div1 type="chapter">
<head type="main">Seasons and Months</head>
<div2 type="section">
<head type="sub">The New Year</head>
<p>
<hi rend="strong">I.</hi> Manx, Yn Vlein Noa. <hi rend
="emphasis">The New Year</hi>. January 1st.
&ldquo;&hellip; anciently the first day of November
was the first day of the year; and the Mummings on the
eve of All Saints&apos; Day still begin their
petition with these remarkable words, &lsquo;To-night
is New Year's night. Og-u-naa. The Moon shines fair
and bright. Tro-la-la.&rsquo;&rdquo; Kelly (<hi
rend="emphasis">circa</hi> 1790), s.v. Baal-saain.
<hi rend="emphasis">See</hi> November 1st.
</p>
</div2>
```

(TEIXLite DTD)