GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

LT resources – lecture 8

Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg
<lars.borin@svenska.gu.se>
<http://spraakbanken.gu.se/personal/lars/>

LT Resources
MLT, 6th October, 2011

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

overview

1. why standards?
2. levels of standards
3. some important standardization efforts
4. why these formats and frameworks?
5. some LT infrastructure initiatives

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

why standards?

Bird & Simons (2003) list the following problem areas for digital language data portability:

1. content
2. format
3. discovery
4. access
5. citation
6. preservation
7. rights

And they conclude:
"We need (. . .) open source tools based on agreed data models (. . .) connected to portable data formats" (Bird & Simons 2003: 580)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

why standards?

It has been argued that attempting standardization for language resources and surrounding information is premature, and the evolving nature of the domain and technology certainly speaks to that claim. But the growth of the web and the explosion in the number of electronic documents to be handled and maintained within the industrial sector has created an immediate and urgent need for generic language processing components for document indexing and classifying, information extraction, summarization, topic detection, etc., in both mono- and multi-lingual environments, together with robust machine translation and facilities for man-machine multimodal communication. While progress will continue, the field has nonetheless reached a point where we can see clear to a reasonable representation and processing model that should fulfill the needs of HLT for at least the foreseeable future. Indeed, commonality that can enable flexible use and reuse of communicative data is essential for the next generation of language processing applications, if we are to build a global information environment. It is therefore critical at this time to move toward standardization, and in particular, to do this in an internationally accepted framework.

(Nancy Ide & Laurent Romary 2007:
Towards international standards for language resources)

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

standards are often overwhelming. . .

Universal Serial Bus Specification Revision 2.0

Table 7-12. Cable Characteristics (Note 14)

Parameter	Symbol	Conditions	Min	Max	Units
VBUS Voltage drop for detachable cables	VBUSD	Section 7.2.2		125	mV
GND Voltage drop (for all cables)	VGNDD	Section 7.2.2		125	mV
Differential Cable Impedance (full-/high-speed)	ZO	(90 Ω ±15%);	76.5	103.5	Ω
Common mode cable impedance (full-/high-speed)	ZCM	(30 Ω ±30%);	21.0	39.0	Ω
Cable Delay (one way)		Section 7.1.16			
Full-/high-speed	TFSCBL			26	ns
Low-speed	TLSCBL			18	ns
Cable Skew	TSKEW	Section 7.1.3		100	ps
Unmated Contact Capacitance	CUC	Section 6.7		2	pF
Cable loss		Specified by table and graph in Section 7.1.17			

Note 1: Measured at A plug.
Note 2: Measured at A receptacle.
Note 3: Measured at B receptacle.

(p. 185 of the USB 2.0 specification [xxviii+622 pp.])

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

. . . but they let you focus on the fun parts

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

levels of standards

wrt the data:

- basic data formats
- metadata standards
- data model frameworks
- information content and structure

wrt status of 'standard':

- widely used formats
- de-facto standards
- international (ISO, W3C, etc.) standards

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

data formats vs. information content

(Why is it not enough to say "It's XML"?)

Eventually, the opener departed for 24, Eoin Morgan taking his second catch in the slips. Wickets continued to tumble, despite the best efforts of Kamran Akmal (27), and Johnston's captaincy was spot on as he brought back Boyd for some extra pace.

The bowler dug a couple in, and both Akmal and Azhar Mahmood spooned catches to Johnston at mid-wicket.

After Mohammad Sami and Iftikhar had added a gutsy 25 for the ninth wicket, spinner McCallan took the last two wickets as wild slogs were held in the deep. Pakistan had been bowled out for 132 in the 46th over.

The rate and manner of vacuum decay are calculated in an explicit flux compactification, including all thick-wall and gravitational effects. For landscapes built of many units of a single flux, the fastest decay is usually to discharge just one unit. By contrast, for landscapes built of a single unit each of many different fluxes, the fastest decay is usually to discharge all the flux at once, which destabilizes the radion and begets a bubble of nothing. By constructing the bubble of nothing as the limit in which ever more flux is removed, we gain new insight into the bubble's appearance. Finally, we describe a new instanton that mediates simultaneous flux tunneling and decompactification. Our model is the thin-brane approximation to six-dimensional Einstein-Maxwell theory.

Somewhere, parently, in the ginnandgo gap between antediluvious and annadominant the copyist must have fled with his scroll. The billy flood rose or an elk charged him or the sultrup worldwright from the excelsissimost empyrean (bolt, in sum) earthspake or the Dannamen gallous banged pan the bliddy du-ran. A scribicide then and there is led off under old's code with some fine covered by six marks or ninepins in metalmen for the sake of his labour's dross while it will be only now and again in our rear of o'er era, as an upshoot of military and civil engagements, that a gynecure was let on to the scuffold for taking that same fine sum covertly by meddlement with the drawers of his neighbour's safe.

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

basic format standards

- e.g.:
  - Unicode <http://www.unicode.org> (≡ ISO/IEC 10646)
  - XML (eXtensible Markup Language) <http://www.w3c.org/XML/> (W3C)
  - JSON (Javascript Object Notation) <http://www.json.org/>
  - RDF (Resource Description Framework) <http://www.w3.org/RDF/> (W3C)
  - . . .
- (plus media formats. . .)

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

metadata standards

- e.g.:
  - Dublin Core <http://dublincore.org/>
  - ISO 639(-1–3) <http://www.sil.org/iso639-3/>
  - OLAC (Open Language Archives Community) <http://www.language-archives.org/>
  - CMDI (Component Metadata Infrastructure) <http://www.clarin.eu/cmdi>

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

data model frameworks

- e.g.:
  - the ISO TC37/SC4 standards and standard proposals <http://www.tc37sc4.org/>
  - the TEI/(X)CES text/corpus formats <http://www.tei-c.org/> <http://www.xces.org/>
  - OWL (Web Ontology Language) <http://www.w3.org/2004/OWL/>

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

information content and structure

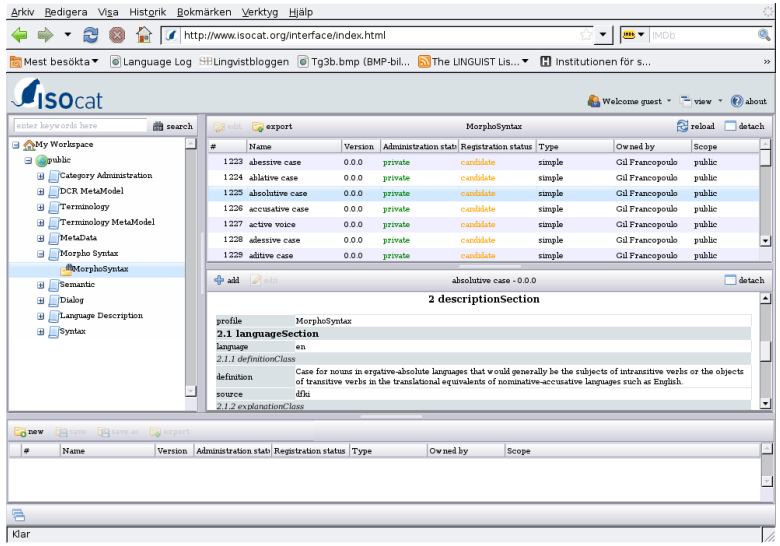
- e.g.:
  - the EAGLES morphosyntactic tagset standard
  - GOLD (General Ontology for Linguistic Description) <http://www.linguistics-ontology.org/>
  - ISO TC37/SC4 DCR <http://www.isocat.org/>

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

TC37/SC4: Data Category Registry (DCR)

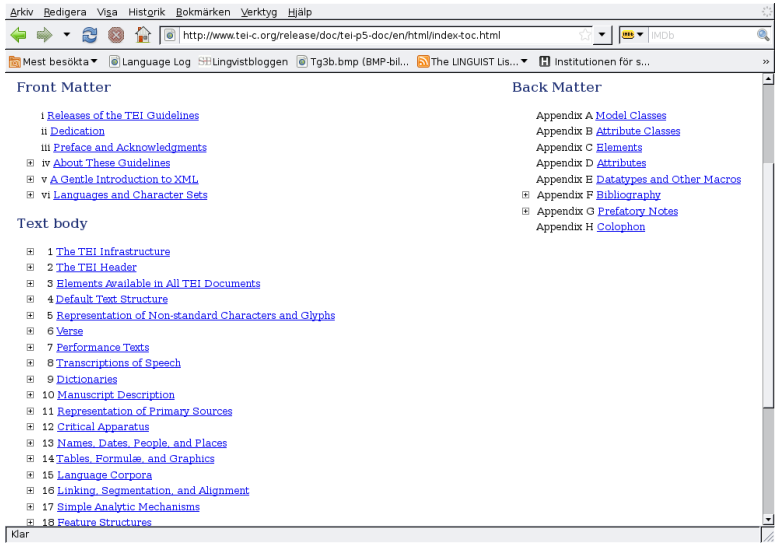


GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

TEI

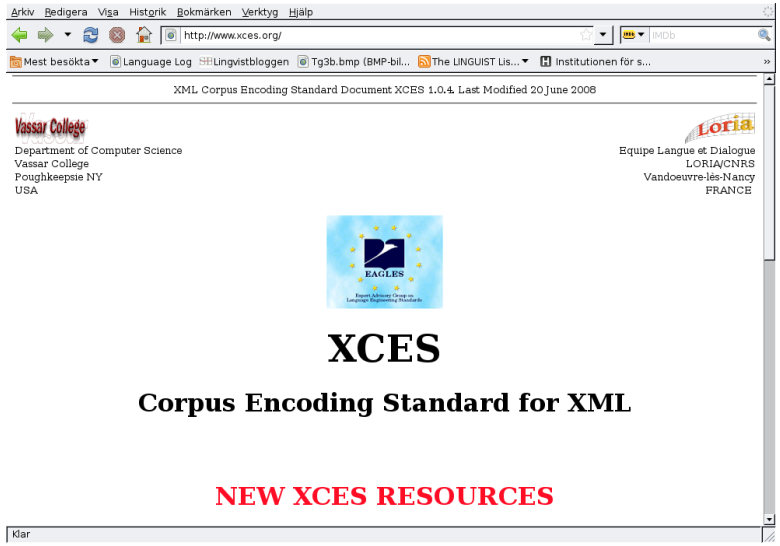


GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

(X)CES



GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

TC37/SC4: Feature structures

INTERNATIONAL STANDARD **ISO 24610-1**

FIrst edition
2006-04-15

Language resource management — Feature structures —

Part 1:
Feature structure representation

Gestion des ressources linguistiques — Structures de traits —

Partie 1: Représentation de structures de traits

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

TC37/SC4: Lexical Markup Framework

INTERNATIONAL STANDARD **ISO 24613**

First edition
2008-11-15

Language resource management — Lexical markup framework (LMF)

Gestion de ressources langagières — Cadre de balisage lexical (LMF)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

some other ISO TC37/SC4 standards

- word segmentation
- linguistic annotation framework
- morpho-syntactic annotation framework
- syntactic annotation framework
- semantic annotation framework

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

OWL

W3C

OWL Web Ontology Language
Overview

W3C Recommendation 10 February 2004

This version:
http://www.w3.org/TR/2004/REC-owl-features-20040210/
Latest version:
http://www.w3.org/TR/owl-features/
Previous version:
http://www.w3.org/TR/2003/PR-owl-features-20031215/
Editors:
Deborah L. McGuinness (Knowledge Systems Laboratory, Stanford University) dlm@ksl.stanford.edu
Frank van Harmelen (Vrije Universiteit, Amsterdam) Frank.van.Harmelen@cs.vu.nl

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

why these formats and frameworks?

First, there are good *extrinsic* reasons for using them:

- Unicode (an ISO standard) and XML (a W3C standard) have a large and growing tool base and can be expected – because they are international standards – to be both portable and long-lived
- There is excellent support for converting into and out of XML from/to other data formats

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

why these formats and frameworks?, 2

There are also some *intrinsic* reasons for using them:

- Unicode and XML are basic; the others build on them, directly or indirectly:
  - RDF (a W3C standard) and the ISO TC37/SC4 standards/ proposals have XML bindings (i.e., agreed ways of expressing them using XML)
  - OWL (a W3C standard) is expressed using RDF
  - metadata are expressed either directly in XML or in RDF/OWL

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

why OWL as linguistic description framework?

- A modern W3C format, OWL has a growing tool base
- OWL promises principled and sound automatic merging of information from distributed information sources (in fact, this is its *raison d'être*)
- Such information sources – called *ontologies* in the terminology of the OWL community – grow in number, in all domains
- There is great interest in providing OWL versions of several ISO TC37/SC4 standards
- Choosing OWL as linguistic representation framework can potentially ensure interoperability between language resources in the narrower sense, multimedia components, and domain knowledge bases

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

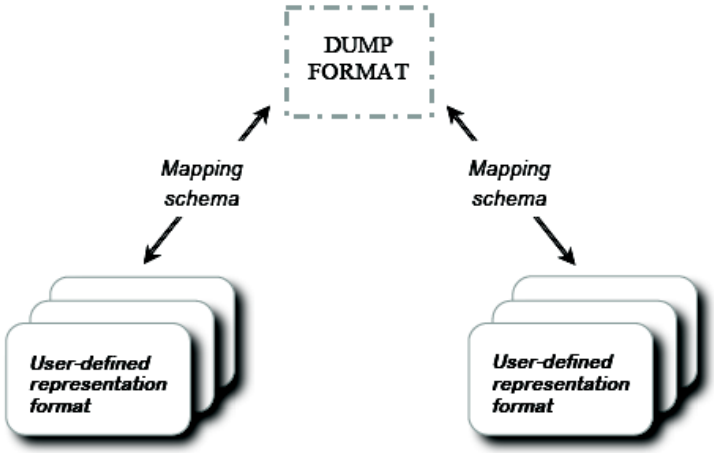
why these formats and frameworks?, 3

Summing up, we could say that

- Unicode provides a standardized way of *expressing* (written) linguistic content
- XML provides a standardized way of *processing form*
- OWL provides a standardized way of *processing content*

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

when in Rome. . .



(Nancy Ide & Laurent Romary 2007:
Towards international standards for language resources)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

some LT infrastructure initiatives

- ELRA <http://www.elra.info>
- LDC <http://www.ldc.upenn.edu>
- CLARIN <http://www.clarin.eu>
- FlareNet <http://www.flarenet.eu>
- META-NET <http://www.meta-net.eu> (META-NORD <http://www.meta-nord.eu>, CAESAR, METANET4U)
- SILT <http://www.anc.org/SILT>
- Språkbanken <http://spraakbanken.gu.se>

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

SILT

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

META-NORD

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Språkbanken