Stabilizing knowledge through standards a perspective for the humanities

Laurent Romary

Stabilizing knowledge through standards

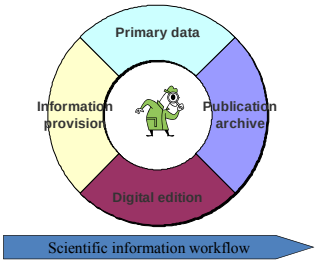
a perspective for the humanities

Laurent Romary
INRIA Gemo & Humboldt Univ. Berlin IDSL

Overview

- From scientific data to lexical databases
- Standardization — TEI, ISO, etc.
- Masculine, feminine, etc.
- Research infrastructures, libraries, etc.

The Scientist's (digital) ecology



Working with research data

- Wide variety and complexity
  - High Energy Physics
    - Particle accelerators / colliders
  - Meteorology
    - Computer simulations
  - Astrophysics
    - Observations, stellar object descriptions
  - Biology
    - Spectrographic representations
  - Linguistics
    - Corpora, grammars, lexical databases

"modern" dictionaries

Petit Larousse, 1905

Simple aims:
- Online rendering
- Precise queries on all fields
- Cross-reference with other dictionaries (dictionnaire de l'Académie)

(source. H. Manuelian, Métadif)

"old" dictionaries

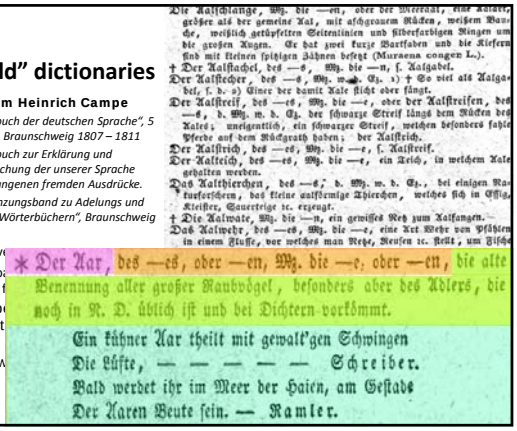
Joachim Heinrich Campe

„Wörterbuch der deutschen Sprache", 5 volumes, Braunschweig 1807 – 1811

„Wörterbuch zur Erklärung und Verdeutschung der unserer Sprache aufgedrungenen fremden Ausdrücke. Ein Ergänzungsband zu Adelungs und Campes Wörterbüchern", Braunschweig 1813

Objectiv
• 6000 p
entries: f
• Testb
Project

(Source W



Full-form lexica

Trésor de la Langue Française - Morphalou

- 539 413 inflected forms, 68 075 lemmas
- Natural Language Processing applications

(Source S. Alt, ATILF-CNRS)

```
chat sms,chat
…
chats smp,chat
…
cheik sms,cheikh:cheik
cheikh sms,cheikh:cheik
…
ferme axs sfs,ferme
ferme ip1s ip3s sp1s sp3s im2s,fermer v
fermement h,fermement
ferment ip3p sp3p,fermer v
ferment sms,ferment
…
```

Multext-East lexicon

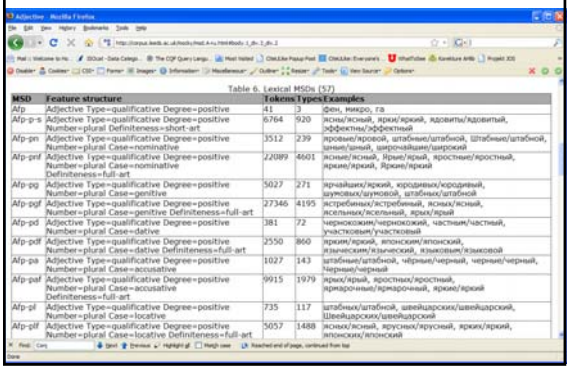
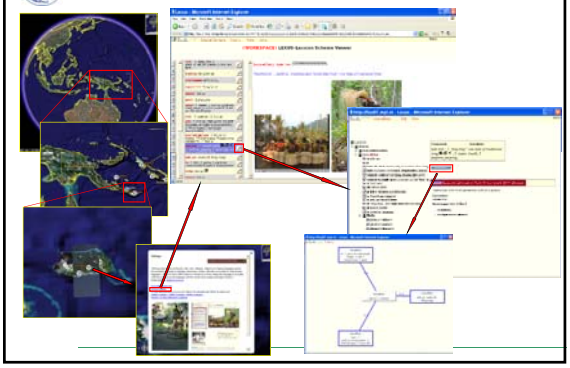


Table 6. Lexical MSDs (57)

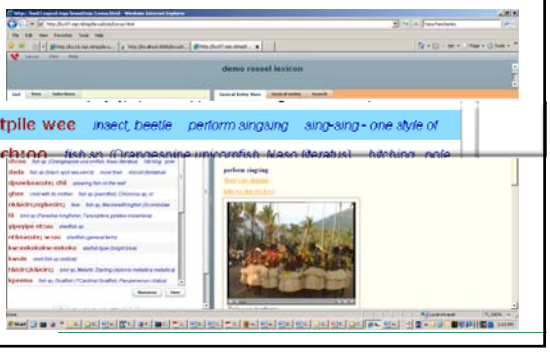
MSD	Feature structure	Tokens	Types	Examples
Afp	Adjective Type=qualificative Degree=positive	41	3	фен, микро, га
Afp-p-s	Adjective Type=qualificative Degree=positive Number=plural Definiteness=short-art	6764	920	ясны/ясный, ярки/яркий, ядовиты/ядовитый, эффектны/эффектный
Afp-pn	Adjective Type=qualificative Degree=positive Number=plural Case=nominative Definiteness=full-art	3512	239	яровые/яровой, штабные/штабной, Штабные/штабной, шины/шиный, широчайшие/широкий
Afp-pnf	Adjective Type=qualificative Degree=positive Number=plural Case=nominative Definiteness=full-art	22089	4601	ясные/ясный, Ярые/яркий, яростные/яростный, яркие/яркий, Яркие/яркий
Afp-pg	Adjective Type=qualificative Degree=positive Number=plural Case=genitive	5027	271	ярчайших/яркий, юродивых/юродивый, шумовых/шумовой, штабных/штабной
Afp-pgf	Adjective Type=qualificative Degree=positive Number=plural Case=genitive Definiteness=full-art	27346	4195	истребимых/истребимый, ясных/ясный, ясельных/ясельный, ярых/яркий
Afp-pd	Adjective Type=qualificative Degree=positive Number=plural Case=dative	381	72	чернокожим/чернокожий, частным/частный, участковым/участковый
Afp-pdf	Adjective Type=qualificative Degree=positive Number=plural Case=dative Definiteness=full-art	2550	860	ярким/яркий, японским/японский, языческим/языческий, языковым/языковой
Afp-pa	Adjective Type=qualificative Degree=positive Number=plural Case=accusative	1027	143	штабные/штабной, чёрные/черный, черные/черный, Черные/черный
Afp-paf	Adjective Type=qualificative Degree=positive Number=plural Case=accusative Definiteness=full-art	9915	1979	яркх/яркий, яростных/яростный, ярмарочные/ярмарочный, яркие/яркий
Afp-pl	Adjective Type=qualificative Degree=positive Number=plural Case=locative	735	117	штабных/штабной, швейцарских/швейцарский, Швейцарских/швейцарский
Afp-plf	Adjective Type=qualificative Degree=positive Number=plural Case=locative Definiteness=full-art	5057	1488	японских/японский, яростных/яростный, ярких/яркий, японских/японский

Jezikoslovno označevanje slovenščine http://nl.ijs.si/jos

Memory of endangered languages



Multimodal lexical information



Why standardizing all this?

- Defining methods or models to facilitate
  - Exchange of data
  - Pooling data from various origins
  - Interoperability between software components
  - Comparability of results
- Involves
  - From a scientific and technological point of view
    - Stabilizing/documenting existing practices, knowledge
    - Looking ahead for potential roadblocks (generalizations)
  - From an organizational point of view
    - International consensus, long term availability and maintenance

Standards: a complex picture

- Standardization bodies or consortia
  - National: AFNOR, ANSI, BSI, DIN, MSA, SIS (Swedish Standard Institute)
  - International: ISO, IEC, CEN, W3C, OASIS, TEI
- Specific fora
  - Many! e.g.
    - LISA (Localization Industry Standards Association)
- Projects with a pre-normative purpose
  - e.g. in Europe:
    - EAGLES, Multext, MATE, ISLE, Lirics, Kyoto

Can scientists bear standards?

- Standards are essentially "bad" for scientists
  - Freezing knowledge
  - Lost of time (which could be dedicated to research)
  - Forcing diverging views to agree
  - …especially if the work is done by others
    [also known as NIH syndrome: "not invented here"]
  - Forcing one to make data readable by others
  - …

How to answer reluctance?

- Main issues
  - Managing the trade-off between *interoperability* and *variability* of linguistic representation
  - Documenting and maintaining document formats
  - Unifying the management, query and presentation of linguistic resources
- A possible answer
  - Standards as specification platforms
- Major factors
  - Expressing constraints on models, adaptation to use cases
  - Identifying generic structures, preventing representation silos

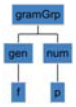
Standardization for language resources: current state

- TEI
  - Initiated in 1987, driving force behind XML creation
  - P5 edition of the guidelines
    - Cf. specification platform (ODD)
- ISO
  - ISO/TC 37: Terminology and language resources
    - ISO/ TC 37/SC 2: ISO 639 series (language codes)
    - ISO/TC 37/SC 3: ISO 16642 (Terminology)
    - ISO/TC 37/SC 4: Language resource management (2002)
- W3C
  - ITS (Internationalization (I18n) activity)
  - SMIL Text (Synchronized Multimedia Integration Language)

Intermezzo — an XML tutorial

- XML is about awful angle brackets

```
<gramGrp>
   <gen>f</gen>
   <num>p</num>
</gramGrp>
```



- XML is about trees
- Issues
  - Specifying structures
  - Providing semantics

Modeling Lexical Structures with the TEI

How it all started

1. Novembre 1987: Vassar College, Poughkeepsie

TEI example

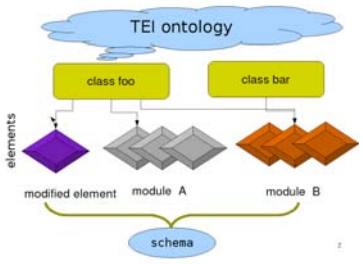
```
<stage>Enter Barnardo and Francisco, two Sentinels,at several doors</stage>
<sp who="Barnardo">
  <l part="f">Who's there?</l>
</sp>
<sp who="Francisco">
  <l>Nay, answer me. Stand and unfold yourself.</l>
</sp>
<sp who="Barnardo">
  <l part="i">Long live the king!</l>
</sp>
<sp who="Francisco">
  <l part="m">Barnardo?</l>
</sp>
<sp who="Barnardo">
  <l part="f">He. </l>
</sp>
```


Following the TEI spirit

Conformance to the TEI means:

- Sharing a common text encoding culture
- Sharing the same vocabulary (when applicable)
- Allowing user autonomy in defining modifications (extensions, customization), but sharing the mechanisms to do so

TEI architecture — playing Lego



List of TEI Modules

	Module name	A short description
add	analysis	Simple analytic mechanisms
add	certainty	Certainty and uncertainty
add	core	Elements common to all TEI documents
add	corpus	Corpus texts
add	dictionaries	Printed dictionaries
add	drama	Performance texts
add	figures	Tables, formulæ, and figures
add	gaiji	Character and glyph documentation
add	header	The TEI Header
add	iso-fs	Feature structures
add	linking	Linking, segmentation and alignment
add	msdescription	Manuscript Description
add	namesdates	Names and dates
add	nets	Graphs, networks, and trees
add	spoken	Transcribed Speech
add	tagdocs	Documentation of TEI modules
add	textcrit	Critical Apparatus
add	textstructure	Default text structure
add	transcr	Transcription of primary sources
add	verse	Verse structures

Encoding a dictionary entry

```
<entry>
   <form>
      <orth>table</orth>
   </form>
   <def>Pièce de mobilier…
   <cit>
      <quote>Une table de cuisine</quote>
   </cit>
</entry>
```

Selecting content
<pos>, <gen>, <num>, <tense>

Constraining content
f., f, fem, feminin, feminine,…

Adding content
e.g.: <transitivity>

Inflectional variants

✱ Der Aar, des —es, oder —en, Mz. die —e, oder —en, die alte Benennung aller großer Raubvögel, besonders aber des Adlers, die noch in N. D. üblich ist und bei Dichtern vorkömmt.

Der Aar, des –es, oder –en,

```
<form type="inflected">
   <gramGrp>
      <case>genitive</case>
      <number>singular</number>
   </gramGrp>
   <form type="determiner">
      <orth>des</orth>
   </form>
   <form type="headword">
      <orth>
         <oVar><oRef/>-es</oVar>
      </orth>
   </form>
   ...
</form>
```

Folie 25

Specification and documentation

TEI's literate programming with ODD (One Document Does it all) provides: schema specification (DTD, RelaxNG, W3C), user oriented documentation, modularity, classes, extensibility.

Before we go any further…

- Which normative reference for the values of element like <gen> (grammatical gender)?
  - Not an issue specific to dictionary design
    - Cf. linguistic annotation at large (e.g. POS tagging)
  - Not an issue specific to the TEI community
    - Such values and their semantics should be defined independantly of any specific tagset
- Is <gen> a self-standing notion?

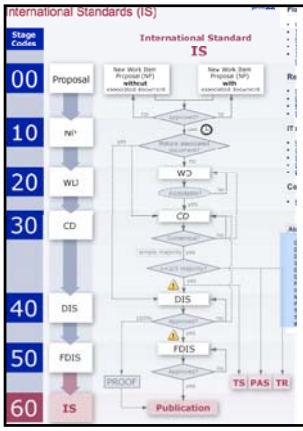
Modeling Lexical Resources within ISO/TC 37/SC 4

ISO in short

- International Organization for Standardization (http://www.iso.org)
  - Administrative view
    - Federation of national standardization bodies
  - Technical view
    - Organized in technical committee and sub-committees
      - ISO technical committees

ISO: a standardisation body

- Providing unique references
  - Language (ISO 639), country (ISO 3166) and script coding (ISO 15924)
    - zh-SG (Chinese for Singapore)
    - sr-Cyrl (Serbian written with Cyrillic script)
- Providing definitions and principles
  - Character encoding
    - ISO 636, ISO 8859-x, ISO 10646/Unicode
- Standard as an evolving material



ISO process

CD = Committee Draft
DIS = Draft International Standard
DPAS = Draft Publicly Available Specification
DTR = Draft Technical Report
DTS = Draft Technical Specification
FDIS = Final Draft International Standard
IS = International Standard
NP = New Work Item Proposal
PAS = Publicly Available Specification
TR = Technical Report
TS = Technical Specification
WD = Working Draft

ISO/TC 37/SC 4 projects

ISO 24610-1:2006 Feature structures -- Part 1: Feature structure representation **ISO/DIS 24610-2 Feature structures -- Part 2: Feature system declaration** **ISO 24613:2008 Lexical markup framework (LMF)**	**ISO/CD 24612 Linguistic annotation framework (LAF)** **ISO/NP 24619 Citation of Electronic Resources (CitER)** **ISO/WD 24616 Multi lingual information framework (MLIF)** **ISO/DIS 24611 Morpho syntactic annotation framework (MAF)** **ISO/CD 24615 Syntactic annotation framework (SynAF)**
ISO/DIS 24617-1 Semantic annotation framework (SemAF) -- Part 1: Time and events **ISO/CD 24617-2 Semantic annotation framework (SemAF) -- Part 2: Dialogue acts**	**ISO/CD 24614-1 Word segmentation of written texts for mono-lingual and multi-lingual information processing -- Part 1: General principles and methods** **ISO/WD 24614-2 Word segmentation of written texts for mono-lingual and multi-lingual information processing -- Part 2: Word segmentation for Chinese, Japanese and Korean**

General modeling framework

- Meta-model
  - General, underlying model that informs current practice
- Data-categories
  - Provides the elementary descriptors to instantiate models

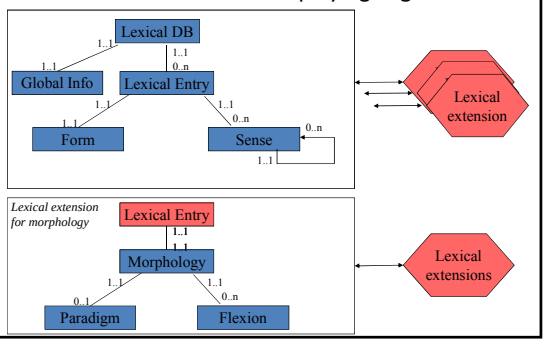
Application to lexical structures

LMF — Lexical Markup Framework
(ISO 24613)

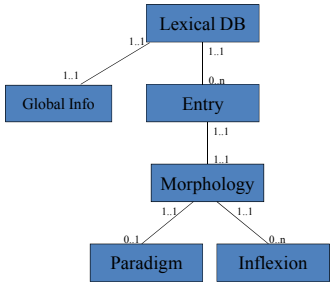
LMF as an ISO project

- Summer 2003: new work item proposal (US) delegation
- Fall 2003: technical proposal (FR) for a data model dedicated to NLP lexica
- ISO 24613
  Convenor:
  - Nicoletta Calzolari (IT)
  - Editors:
    - Gil Francopoulo (FR), Monte George (US)
  - 13 versions written, dispatched (to the National delegations nominated experts), commented and discussed in various ISO technical meetings
- IS (= published standard) in oct. 2008

LMF architecture — playing Lego

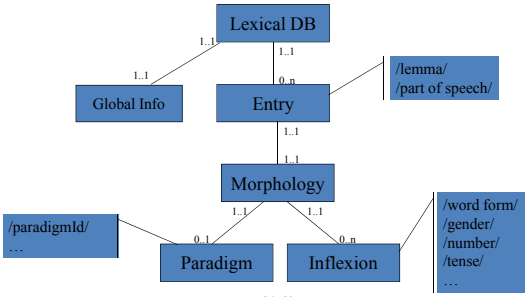


Example: designing a full-form lexicon



Seite 37

Decorating the model



Seite 38

A possible XML instance

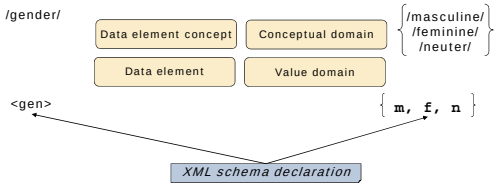
```
<lexicalEntry>
   <lemma>chat</lemma>
   <grammaticalCategory>noun</grammaticalCategory>
   <morphology>
    <paradigm>
        <paradigmIdentifier>fr-s-plural</paradigmIdentifier>
    </paradigm>
    <inflexion>
        <wordForm>chat</wordForm>
        <number>singular</number>
    </inflexion>
    <inflexion>
        <wordForm>chats</wordForm>
        <number>plural</number>
    </inflexion>
    …
   </morphology>
</lexicalEntry>
```

Seite 39

A central concept: data category

- Definition
  - Elementary descriptor used in a linguistic description or annotation scheme
- Examples
  - Fields: /part of speech/, /grammatical gender/
  - Values: /feminine/, /plural/, /dual/, /ablative case/
- Role
  - Specification
  - Documentation
- A reference space for schema designers
  - Towards an international registry for language resources
    - Data Category Registry (DCR); cf. ISO 12620

Formal background: ISO 11179




```
<gen>f</gen>
```

Some deeper thoughts on *gender*

- A central category in linguistic and computational linguistic
  - Lexica, morpho-syntactic tagging, agreement in syntax, etc.
- Can we standardize “gender”
  - Interoperability vs. language variety
- By the way, gender is not exactly “sex”
  - ISO 5218, Information technology — Codes for the representation of human sexes
    - 0 = not known; 1 = male; 2 = female; 9 = not applicable

The linguistic view

- What is gender:
  - "a classification of nominals, as shown by agreement"
    - E.g. die Katze – der Hund
  - Determiners, adjectives, numerals, verbs
    - E.g. Control by anaphoric pronouns (cf. en)
      - Die Katze… sie…
  - Not present in all languages
- Number of genders (Greville G. Corbett)

Application: Independent personal pronouns

- Example: Rif Berber (McClelland 2000: 27)

1sg	nəš	1pl	nəšnin
2sg.m	š ə k	2pl.m	k ə niw
2sg.f	šəm	2pl.f	kanint
3sg.m	nətta	3pl.m	nitnin
3sg.f	nəttæθ	3pl.f	nitənti

- Gender Disctinctions in Independent Personal Pronouns, Source: Anna Siewierska (cf. wals.info)

The TC 37 model — ISO 12620

Entry Identifier:	grammatical gender
Profile:	morpho-syntax
Definition (fr):	Catégorie grammaticale reposant, selon les langues et les systèmes, sur la distinction naturelle entre les sexes ou sur des critères formels (Source: TLFi)

Object Language: fr *Name*: genre *Conceptual Domain*: {/feminine/, /masculine/}	*Object Language*: en *Name*: gender *Name*: grammatical gender	*Object Language*: de *Name*: Geschlecht *Name*: Genus *Conceptual Domain*: {/feminine/, /masculine/, /neuter/}

Convergence?

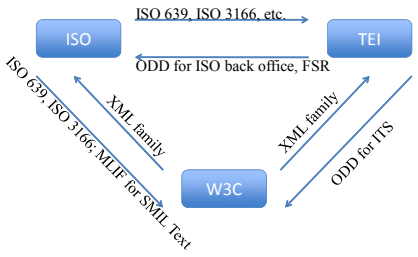
Petit Larousse 1905 by Métadif (source H. Manuélian) ➔ goes TEI

Campe by Uni. Würzburg (W. Wegstein) ➔ goes TEI

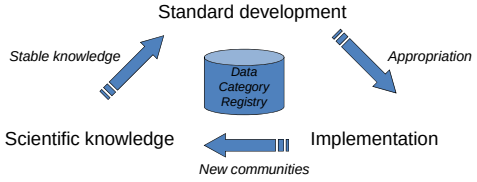
Morphalou 2005 by ATILF (S. Alt) ➔ goes LMF

```
<lexicalEntry>
  <lemma>cheikh</lemma>
  <spellingVariant>cheik</spellingVariant>
  <grammaticalCategor>common noun</grammaticalCategory>
  <grammaticalGender>masculine</grammaticalGender>
<morphology>
  <inflexion>
    <wordForm>cheikh</wordForm>
    <grammaticalNumber>singular</grammaticalNumber>
  </inflexion>
  <inflexion>
    <wordForm>cheikhs</wordForm>
    <grammaticalNumber>plural</grammaticalNumber>
  </inflexion>
        </morphology>
</lexicalEntry>
```

Convergence?



Standards as an emanation from scientific knowledge



Epilogue

RESEARCH INFRASTRUCTURES IN THE HUMANITIES

Research Infrastructures

- In general: permanent and physical
- Natural sciences: ice breakers for polar research, satellites, telescopes, particle accelerators, laboratories
- RIs for the humanities?
  - **Cultural heritage** in all forms is the main source of humanities research
  - **Libraries and archives** are the traditional "laboratories" for the humanities
- In the digital age, essential for innovative humanities research is:
  - Access to digitised heritage data (data bases, text corpora, speech, image collections, etc.)
  - Tools to process this information

Core activities

- Digitise – Curate – Preserve
  - Standards development and promotion
  - Curation, preservation and digitisation services
  - Technology platforms
  - Legal services and advice
- Discover – Access – Deliver
  - Authentication and authorisation,
  - Harvesting, aggregating, hosting
  - User-friendly discovery, delivery and use
- Connect – Collaborate – Use
  - Supporting communities of practice
  - Facilitating new research practice
  - Tools and registries

(conclusive) priorities

- Mastering the technology
  - Not all scientist are technological geeks
    - Transparency
- Answering priority needs
  - Strong request to provide infrastructures for simple types of data
    - Pragmatic sense
- Preserving scientific patrimony
  - High amounts of research data is continuously lost
    - Identification, preservation

Should we/you be afraid of standards?

```
<cit>
    <quote>Yes you should be afraid, but you should be
    more afraid of not having them</quote>
    <author>Wendell Piez</author>
</cit>
```