Stabilizing knowledge through standards a perspective for the humanities
Laurent Romary

To cite this version:

HAL Id: inria-00438724
https://hal.inria.fr/inria-00438724
Submitted on 4 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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. Marzulli, Hetad)

“old” dictionaries
Joachim Heinrich Campe
Wörterbuch der deutschen Sprache, 3 volumes, Braunschweig 1807–1811
Wörterbuch zur Erläuterung und Verdeutschung der unserer Sprache aufgeprägten fremden Ausdrücke
Ein Ergänzungsbuch zu Adelungs Wörterbuch der deutschen Sprache, 1813
Objectives:
• 6000 pages
• Testbed for the TextGrid Project
(Source: W. Wegstein, Univ. Würzburg)
Full-form lexica

Tresor de la Langue Française - Morphatou
• 539-413 inflected forms, 68 075 lemmas
• Natural Language Processing applications

(Marc S. Alt, ATILF-CNRS)

chat sme, chat
chat sme, chat
cheikh sme, cheikh: cheikh
ferme sms sms, ferme
ferme ips ips ips ips, fermer v
fermenent h, ferment
ferment sms, ferment

Multext-East lexicon

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
    • Stabilising/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.
    - USA (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
  - CT specification platform (ODD)
- ISO
  - ISO/TC 37: Terminology and language resources
  - ISO/TC 37/SC 2: ISO 639 series (language codes)
- W3C
  - I18n (Internationalization) activity
  - SMIL Text (Synchronized Multimedia Integration Language)

Intermezzo — an XML tutorial

- XML is about awful angle brackets
- XML is about trees
- Issues
  - Specifying structures
  - Providing semantics

Modeling Lexical Structures with the TEI
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

Encoding a dictionary entry

```
<Entry>
  <Form>
    <Orth>table</Orth>
  </Form>
  <GramGrp>
    <Pos>n</Pos>
    <Gen>f.</Gen>
  </GramGrp>
  <Def>Pièce de mobilier…</Def>
  <Cit>
    <Quote>Une table de cuisine</Quote>
  </Cit>
</Entry>
```
Inflectional variants

Der Aar, des –es, oder –en,  

 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 linguistc annotation at large (e.g. POS tagging)  
  – Not an issue specific to the TEI community  
  – Such values and their semantics should be defined independently of any specific tagset

• Is <gen> a self-standing notion?

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
General modeling framework

• Meta-model
  – General, underlying model that informs current practice

• Data-categories
  – Provides the elementary descriptors to instantiate models

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 lexic

• ISO 24613
  Convenor:
  – Nicoletta Calzolari (IT)
  – Editors:
    – Gil Francisco (FR), Monte George (US)
  – 13 versions written, delivered (to the National delegations nominated experts), commented and discussed in various ISO technical meetings.
• IS (= published standard) in oct. 2008

Application to lexical structures

LMF — Lexical Markup Framework (ISO 24613)

LMF architecture — playing Lego
Example: designing a full-form lexicon

Lexical DB

Entry

Global Info

Morphology

Paradigm

Inflexion

Decorating the model

Lexical DB

Entry

Global Info

Morphology

Paradigm

Inflexion

Global Info

Entry

Morphology

Paradigm

Inflexion

A possible XML instance

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

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], [fem], [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

- **Data element concept**
- **Conceptual domain**

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)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>1sg.m</td>
<td>Rif</td>
<td>Rif</td>
</tr>
<tr>
<td>2sg.m</td>
<td>Berber</td>
<td>Berber</td>
</tr>
<tr>
<td>3sg.m</td>
<td>der</td>
<td>der</td>
</tr>
<tr>
<td>1sg.f</td>
<td>Rif</td>
<td>Rif</td>
</tr>
<tr>
<td>2sg.f</td>
<td>Fenera</td>
<td>Fenera</td>
</tr>
<tr>
<td>3sg.f</td>
<td>der</td>
<td>der</td>
</tr>
</tbody>
</table>

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

The TC 37 model — ISO 12620

Convergence?

ISO, ISO 639, ISO 3166, etc. -> TEI

Standards as an emanation from scientific knowledge

- Stable knowledge
- New communities
- Scientific knowledge
- Implementation
- Data Category Registry
- Standard development
- Appropriation
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 scientists 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>