



Complexity and simplicity in eInfrastructures - some issues

Laurent Romary
Max Planck Digital Library

DINI meeting, Oldenburg, 8 Okt. 2007



What I will not speak about

- Long term archiving
- Publications and publication repositories
- Open access
- Grid computing
- eSciDoc... or you will not notice it



Background

- Many initiatives to provide research infrastructures at national and EU level
- What do they or will they offer?
- Do we match the scientists' expectations?
- Access, standards and community building.

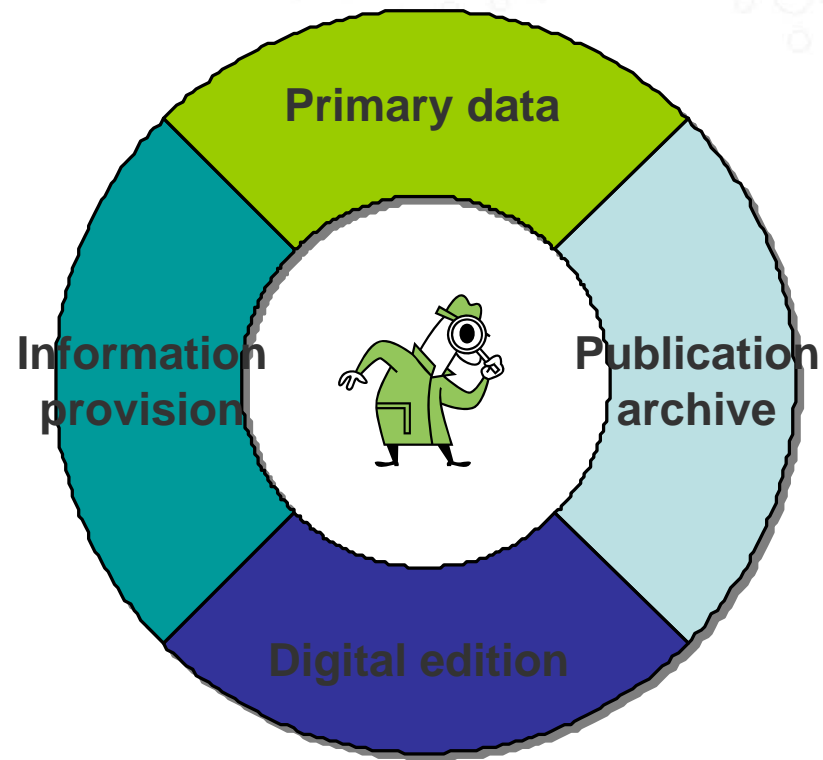
...with a view on the humanities



Why do we need eInfrastructures?

- The scientist's ecology
- Central role of research data and primary sources

The Scientist's (digital) ecology



Scientific information workflow



Working with research data

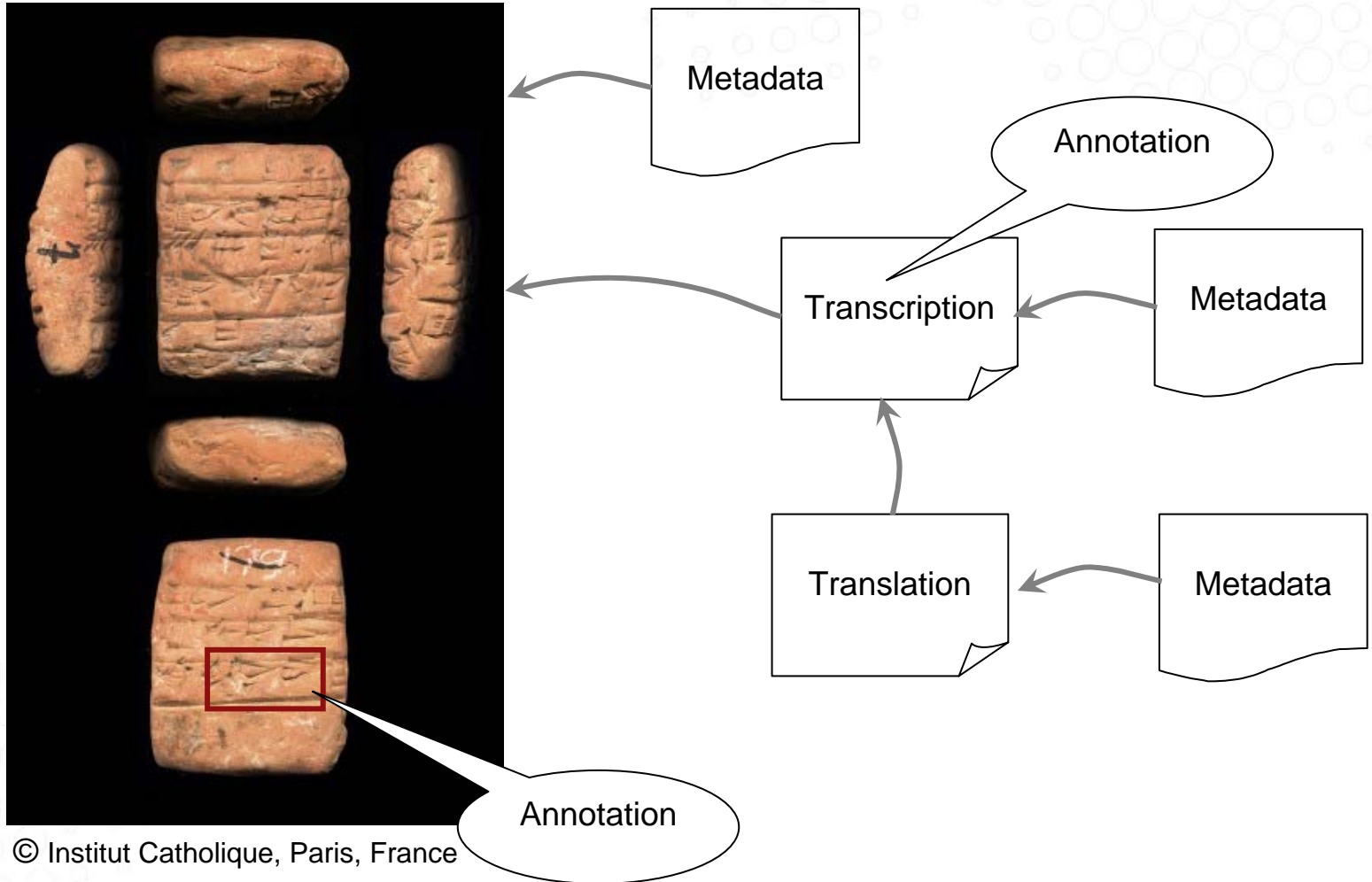
- Wide variety and complexity
 - High Energy Physics
 - Particle accelerators / colliders
 - Meteorology
 - Computer simulations
 - Astrophysics
 - Observations, stellar object descriptions
 - Biology
 - Spectrographic representations
 - Linguistics
 - Corpora, grammars, lexical databases



Research Infrastructures

- RIs in general: **permanent** and **physical**
- RIs for the 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

Working with primary (digital) sources in the humanities





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



Some guidelines

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



Dealing with the complexity

- A selection of core issues
 - Standardisation
 - Tools
 - Documentation
 - Curation

- A specific application
 - Lexical data description



Background: Representing lexical data

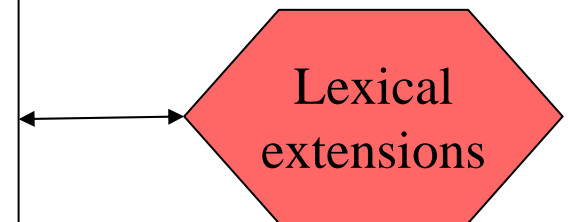
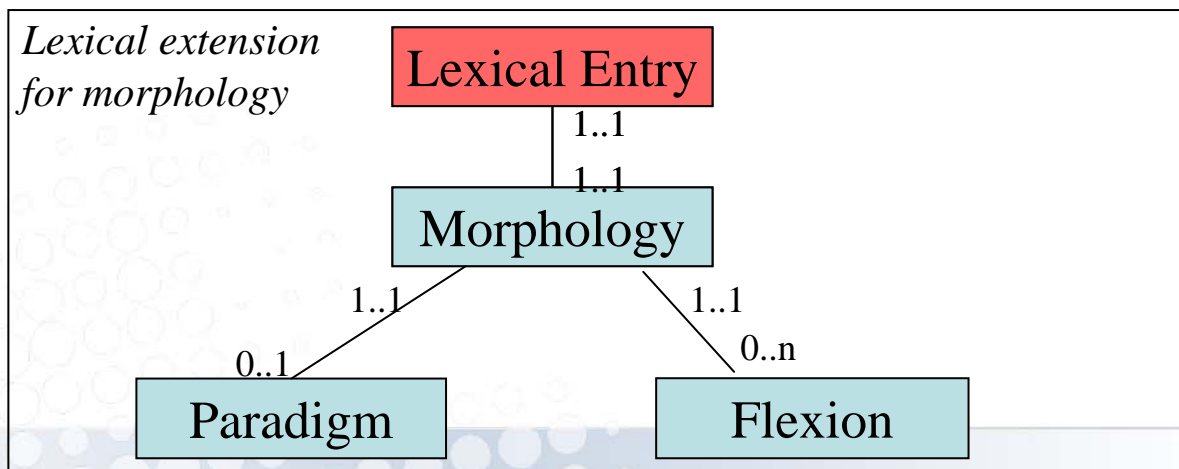
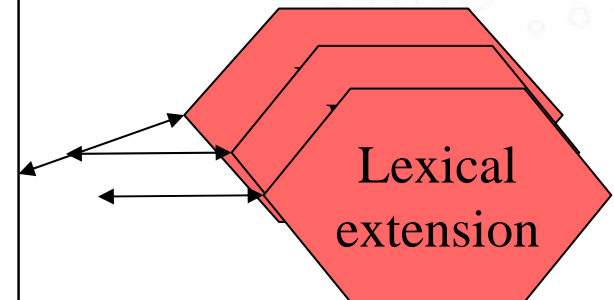
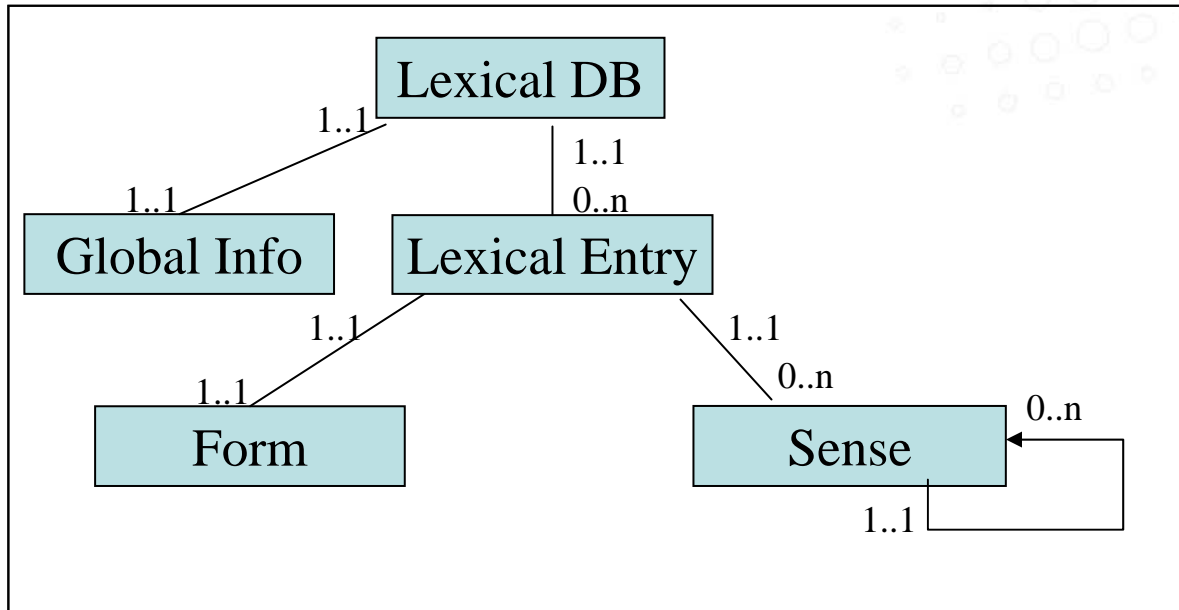
- First level of abstraction in linguistic analysis
 - Psycholinguistic, field linguistics, computational linguistics
- A huge amount of legacy information
 - Proprietary formats
 - Proprietary tools
 - E.g. Shoebox
- Variety of needs
 - Language description, dictionary making, automatic processing



Issue 1: Can scientists bear standards?

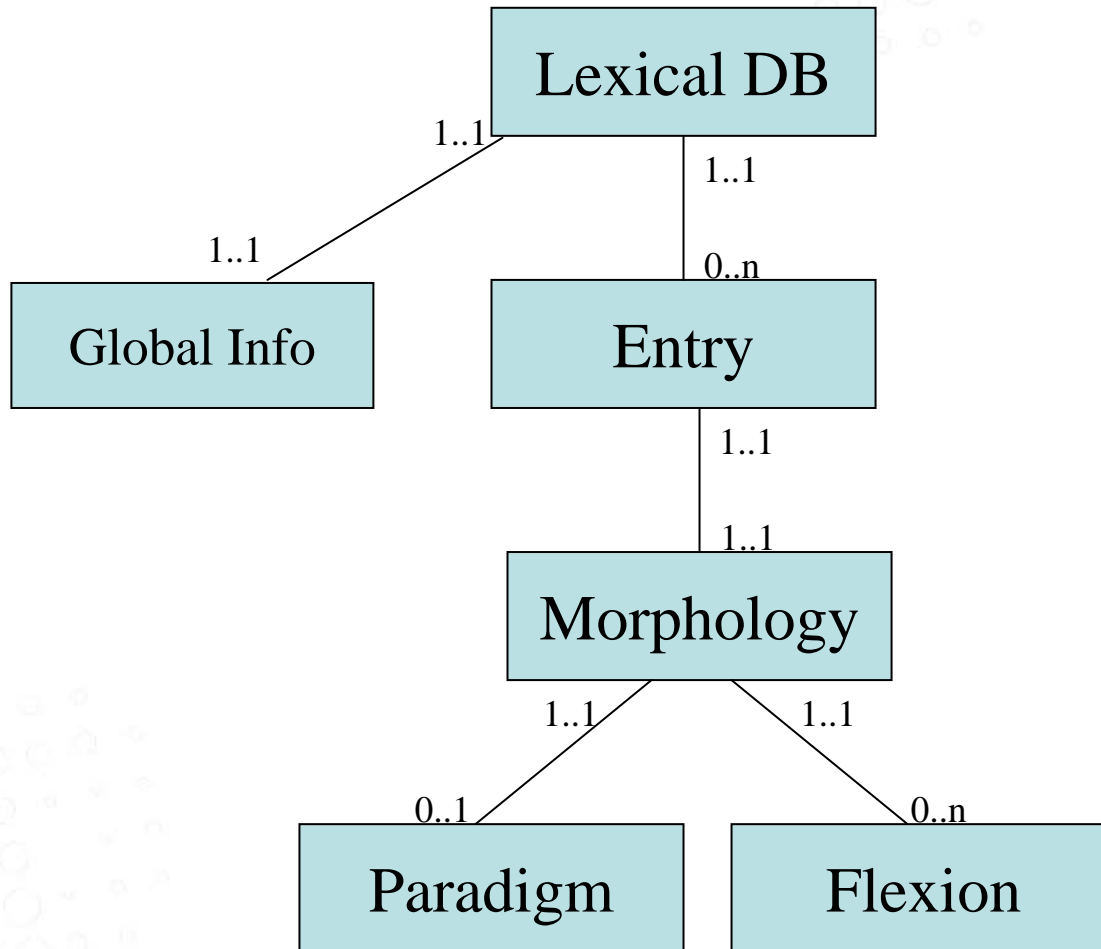
- Standards are essentially bad for scientists
 - Freezing knowledge
 - Making one lose time that could be dedicated to research
 - Forcing diverging views to agree
- A positive view on standards
 - Documenting data
 - Giving semantics to data
 - Pooling data from various origins
 - Allowing interoperability of tools
- A possible answer
 - Standards as specification platforms
 - Example:
 - ISO committee TC 37/SC 4 (Language resource management)
 - LMF - Lexical Markup Framework

ISO 24613 LMF: Lexical Markup Framework



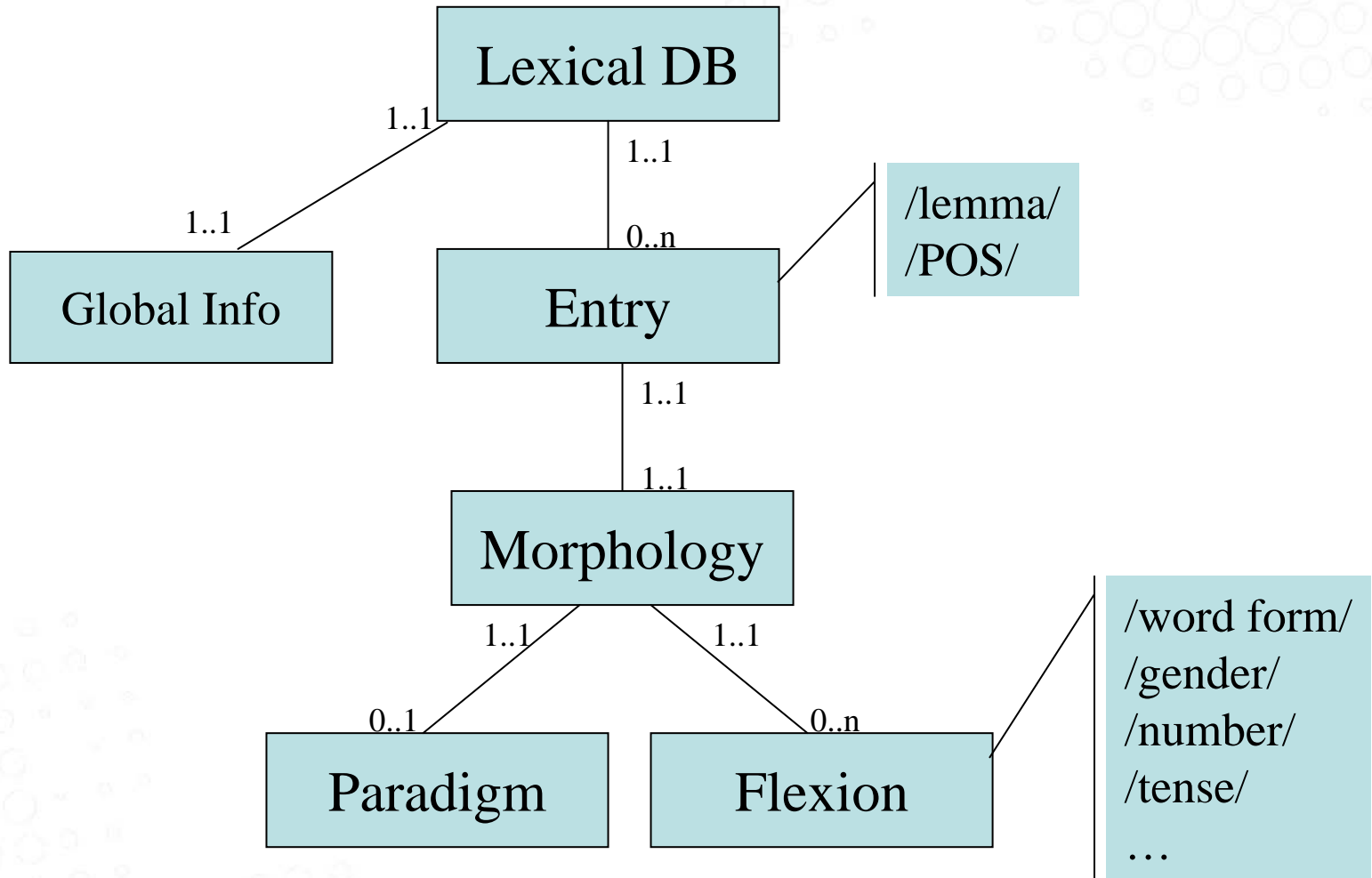


Meta-model for morphological description





Decorating the model





For those who want to see some XML?

```
<struct type='lexical entry'>
  <feat type='lemma'>chat</feat>
  <feat type='grammatical category'>noun</feat>
  <struct type='morphology'>
    <struct type='paradigm'>
      <feat type='paradigm identifier'>fr-s-plural</feat>
    </struct>
    <struct type='inflexion'>
      <feat type='word form'>chat</feat>
      <feat type='number'>singular</feat>
    </struct>
    <struct type='inflexion'>
      <feat type='word form'>chats</feat>
      <feat type='number'>plural</feat>
    </struct>
    ...
  </struct>
</struct>
```

Providing semantics through data categories

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)

Definition (en): Grammatical category... (Source: TLFi (Trad.))

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/}



Issue 2: Tools to make this transparent

- Expected functionalities
 - Data modelling
 - Data management
 - Legacy data recovery

- Example: Lexus
 - On-line lexical data management tool
 - Implemented at MPI for Psycholinguistic, Nijmegen
 - Fully implements LMF
 - Uses TEI/ODD for XML schema specification

Data modelling

LEXUS Lexicon Structure View Search Relation types Help jacrin

english lexicon demo: Lexicon structure

LEXUS

- lexicon
 - lexiconInformation
 - creation date
 - last modification
 - lexicalEntry
 - partOfSpeech
 - headword
 - form
 - pronunciation
 - phoneticForm
 - multi media
 - image
 - video
 - sound
 - inflected form
 - orthography
 - person
 - grammaticalNumber
 - tense
 - sense
 - gloss
 - definition
 - semantic domain
 - synonym
 - example
 - example sentence

Component: lexicalEntry

General
DataCategories
Components

Creation date: 2007-06-12 Last modified: 2007-09-14 by jacrin

Description:
Represents a word, a multi-word expression, or an affix in a given language

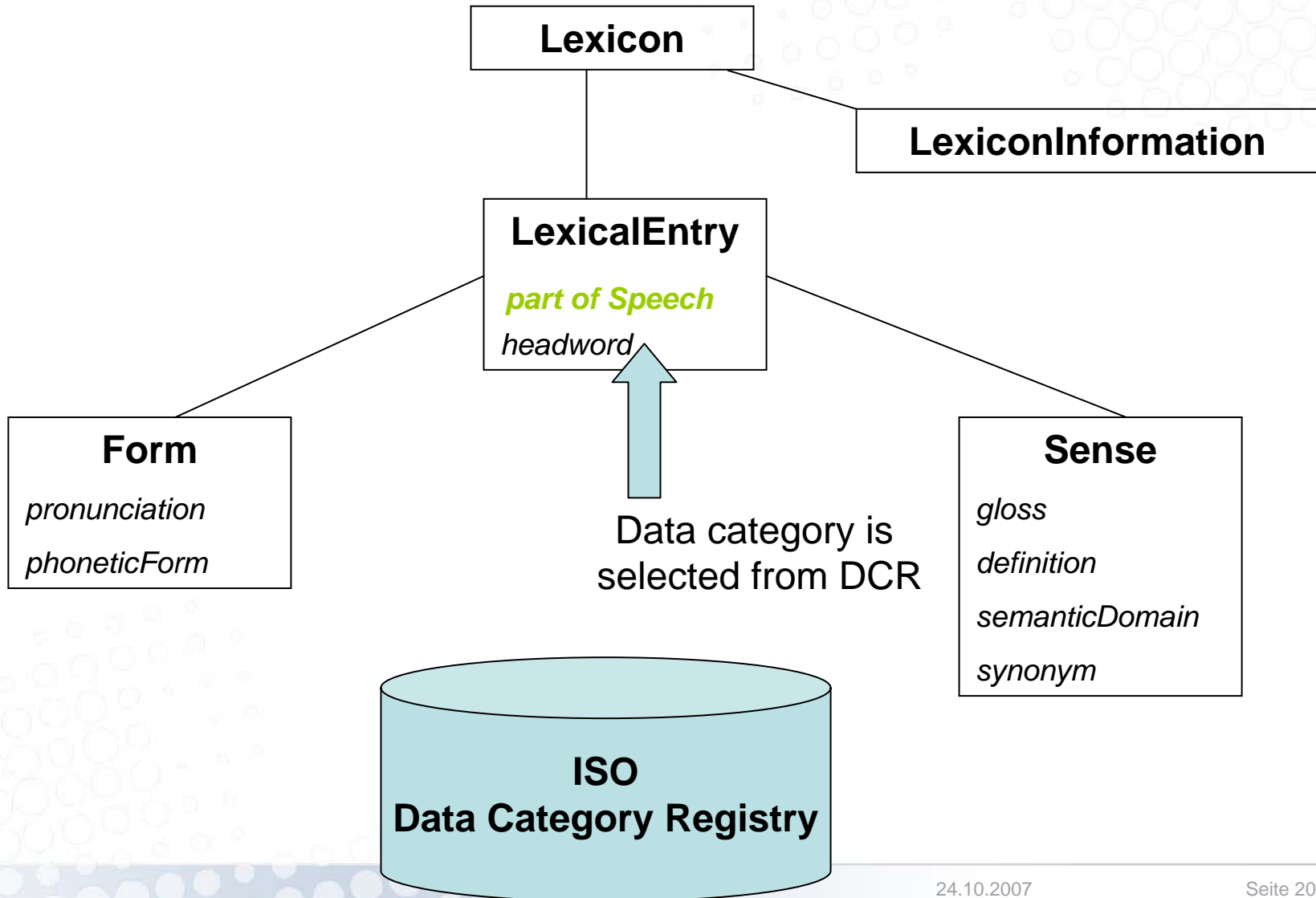
Admin info:
to be filled out

Model name: LMF Core

mandatory

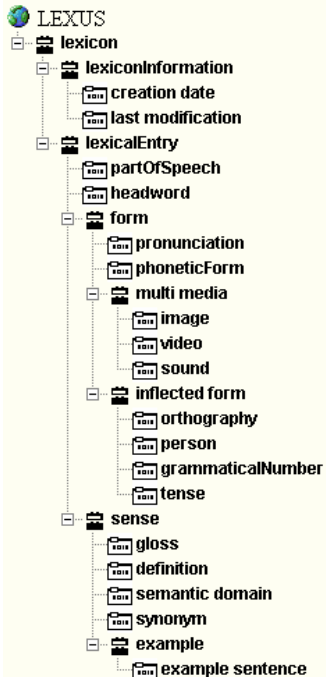
multiple values allowed

Data Category Usage Platform



Access to data categories

english lexicon demo: Lexicon structure



DataCategory details partOfSpeech - Microsoft Internet Explorer provided by MPI Nijmegen

Data category information

partOfSpeech
(MorphoSyntax)

Administration Identification

Identifier: partOfSpeech	Creation date: 2004-07-09
Version: 0.0.0	?
Registration authority: Private	
Registration status: candidate	
Administration status: Private	Last change date: 0000-00-00
Origin: ?	?

Description

Profile: MorphoSyntax

ConceptualDomain: adjective, adposition, adverb, affirmativeParticle, affixedPersonalPronoun,

Integration in the model

LEXUS Lexicon Structure View Search Relation types Help jacrin

english lexicon demo: Lexicon structure

LEXUS

- lexicon
 - lexiconInformation
 - creation date
 - last modification
 - lexicalEntry
 - partOfSpeech
 - headword
 - form
 - pronunciation
 - phoneticForm
 - multi media
 - image
 - video
 - sound
 - inflected form
 - orthography
 - person
 - grammaticalNumber
 - tense
 - sense
 - gloss
 - definition
 - semantic domain
 - synonym
 - example
 - example sentence

partOfSpeech

user defined
 shoebox
 iso 12620

General

Creation date: 2007-09-14 Last modified: 2007-09-14 Author: jacrin

Description:
Term used to describe how a particular word is used in a sentence.

Admin info:
to be filled out

mandatory
 multiple values allowed

Reference: 1345

Sort Order:

Natural
 Other: a(AaÀàÁáÂâÃãÄäÅåÆæ)b(Bb)c(CcÇç)d(DdÐð)e(EeÈèÉéÊêËë)f(Ff)g(Gg)h(Hh)i(IiÍíÎî)j(Jj)k(Kk)l(Ll)m(Mm)n(NnÑñ)o(OoÓóÔôÕõÖöØø)p(Pp)q(Qq)r(Rr)s(Ss)t(Tt)u(UuÛüÜü) v(Vv)w(Ww)x(Xx)y(YyÝý)z(Zz)

Import from other formats(shoebox)

LEXUS Lexicon Search,... Help Administration jacrin

Workspace of jacrin

Please select the type of lexicon you would like to import

- Shoebox
 Clan
 any XML *Note: schema file required.*

Please specify whether you want to use a type file.

- Create new lexicon
 Create new Lexicon no type file..

Shoebox lexicon schema and data import

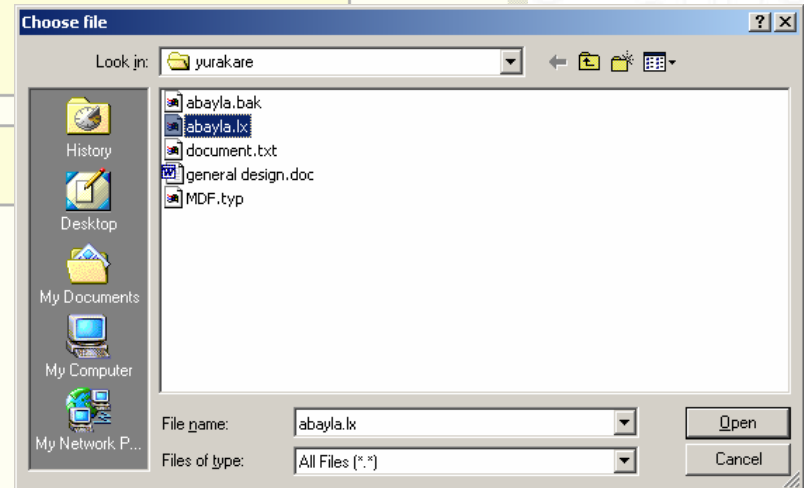
This page allows you to import a shoebox lexicon. A new LexicalDatabase will be created for this lexicon.

Please upload the Shoebox .typ structural file here.

Please upload the Shoebox lexicon file here.

Allow me to restructure the shoebox model.

Please note that the character encoding of your file is expected to be UTF-8 or ISO-LATIN-1.



“Schema” and data files are imported

Import from other formats(shoebox)

LEXUS Lexicon Search... Help Administration jacrin

Workspace of jacrin

marker	name	description	classification
1d	First dual	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
1e	First plural exclusive	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Lexical entry Form
1i	First plural inclusive	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
1p	First plural	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
1s	First singular	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
2d	Second dual	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
2p	Second plural	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
2s	Second singular	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
3d	Third dual	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
3p	Third plural	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
3s	Third singular	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
4d	Non-animate dual	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
4p	Non-animate plural	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
4s	Non-animate singular	Used to give the vernacular for this particular paradigm form. (It is better to use the \pdl field set for this.)	Sense
a	*		Sense
an	Antonym	Used to reference an antonym of the lexeme, but using the \lf (lexical function) field for this is better practice.	Sense
bb	Bibliography	Used to record any bibliographic information pertinent to the lexeme. MDF adds the label 'Read' to this field. For basic information (to reference an entry in a more complete bibliography database) include the following: AuthorLastName, FirstName/Initials, date, pp. For a more complete reference include: AuthorLastName, FirstName/Initials, date, Title, City, Publisher, pp. xxx-yyy.	Sense
bw	Borrowed word (loan)	Used for denoting the source language of a borrowed word.	Sense
bwn	*		Sense
ce	Cross-ref. gloss (E)	Gives the English gloss(es) for the vernacular lexeme referenced by the preceding \lcf field.	Sense

This is a generic reference marker used to link together any two related entries in the

Shoebox schema is remodeled to fit LMF model

Import from other formats(shoebox)

LEXUS Lexicon Structure View Search Relation types Help jacrin

Yurakare example: Lexicon structure

LEXUS

- lexicon
 - lexiconInformation
 - contact person
 - consultant
 - language
 - Grammar Sketch
 - lexicalEntry
 - lexeme
 - part of speech
 - borrowed word
 - reference
 - date
 - New Data Category
 - form
 - phonetic form
 - grammar note
 - picture
 - sound file
 - link to the archive
 - sense
 - english gloss
 - english definition
 - enlish reversal
 - cross reference
 - example
 - vernacular example
 - english translation of example
 - notes
 - ethnographic note
 - sub entry
 - sub entry
 - sub entry sense
 - sub entry english definition
 - sub entry example

part of speech

user defined
 shoebox
 iso 12620

General

Creation date: 2007-09-17 Last modified: 2007-09-17 Author: jacrin

Description:
to be filled out

Admin info:
to be filled out

mandatory

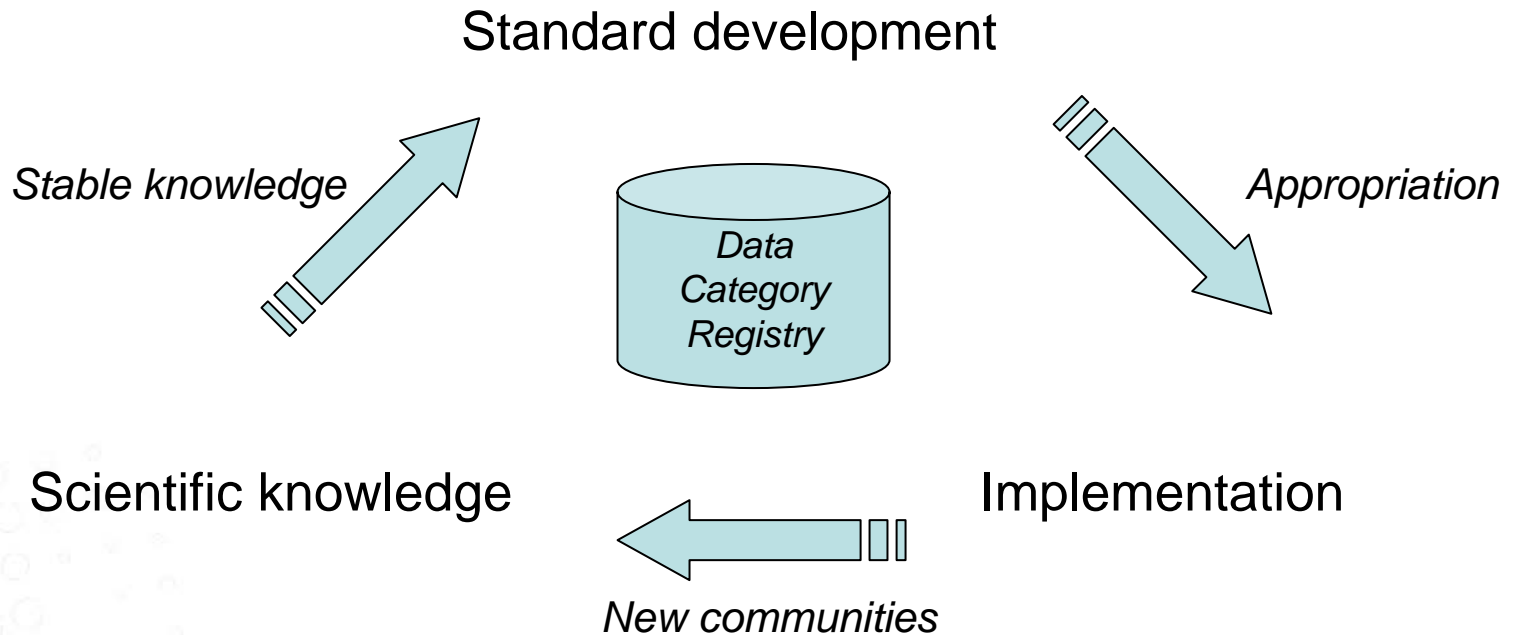
multiple values allowed

Sort Order:

Natural

Other: a(AaÀáÁâÄåÃäÅãÄäÆæ)b(Bb)c(CcÇç)d(DdÐð)e(EeÈéÊêËë) f(Ff)g(Gg)h(Hh)i(IiÍíÎî)j(Jj)k(Kk)l(Ll)m(Mm)n(NnÑñ)o(OoÓóÔôÕõÖöØø)p(Pp)q(Qq)r(Rr)s(Ss)t(Tt)u(UuÛüÜüŮů)v(Vv)w(Ww)x(Xx)y(YyÝÿ)z(Zz)

Standards as an emanation from scientific knowledge





MAX PLANCK
digital library

Gathering expertise — Colab



Issue 3: documenting the complexity

- Standards are necessary, but the picture is complex
 - Legibility of data
 - In space: sharing scientific sources with others
 - In time: pooling together the records of science
 - Generic standards (horizontal)
 - ISO 10646/Unicode, XML, etc.
 - Specific standards
 - ISO-IEC/JTC 1 (MPEG), ISO/TC 37 (ISO 639, TMF), TEI
 - E.g. TEI:
 - A wide range of documented elements for the encoding of textual data
 - A flexible architecture to select the elements adapted to one's needs



MPDL CoLaboratory (MPDL CoLab)

- Platform for community building and knowledge exchange
- Aim:
 - improve exchange of explicit knowledge and make tacit and individual know-how explicit
- Supports community-building processes
 - Connects people with similar fields of interest and goals
 - within the MPS: MPDL, librarians, scientists
 - Outside: underlying basis of our national and international collaborations
- Provide information about existing standards and best practices in the domain of supporting scientific life cycles
 - Ensuring long-term compatibility between local and centralized initiatives within the MPDL

QuickTime™ et un
décompresseur TIFF (non compressé)
sont requis pour visionner cette image.



Issue 4: do we need libraries, or librarians

- Library as a place
 - Core reference monographies
 - Complementarity with centralized archives
 - Local management of primary sources
 - Selection, digitization, access
 - Library as digital curation centres
 - Centre of gravity of scientific information (cf. Bibliothek2007)
- Library as function
 - From information provision to information management
 - Identification of a “digital curator” profile: interface between scientists and scientific information
 - Local mirrors of central activities
 - We probably do need even more librarians...



Final words

- e-Infrastructures
 - We need them => which model fits which scientific community
- Communities
 - Sharing content and practices
- Helping scientist document the semantic of their data
 - Central role of data category registries
- Central-decentral (is it just a question for the MPS?)
 - A constant balance between contradictory forces
- Objective: simplicity for scientists