

Towards a new Generation of Language Industry Standards – the contribution of ISO/TC 37

2007-07-09

OFMR 2007, NYC

Gerhard Budin
University of Vienna
Center for Translation Studies

Chair, ISO TC 37/SC 2

Overview

1. Terminology Management as an Instrument of Standardization Management
2. Principles and Methods of Terminology Standardization
3. Term Formation in International Terminology Standardization
4. Standards for Terminology Management
5. ISO/TC 37 standards and scenarios for their application
6. Language Industry Standards
7. Conclusions

Terminology Management as an Instrument of Standardization Management

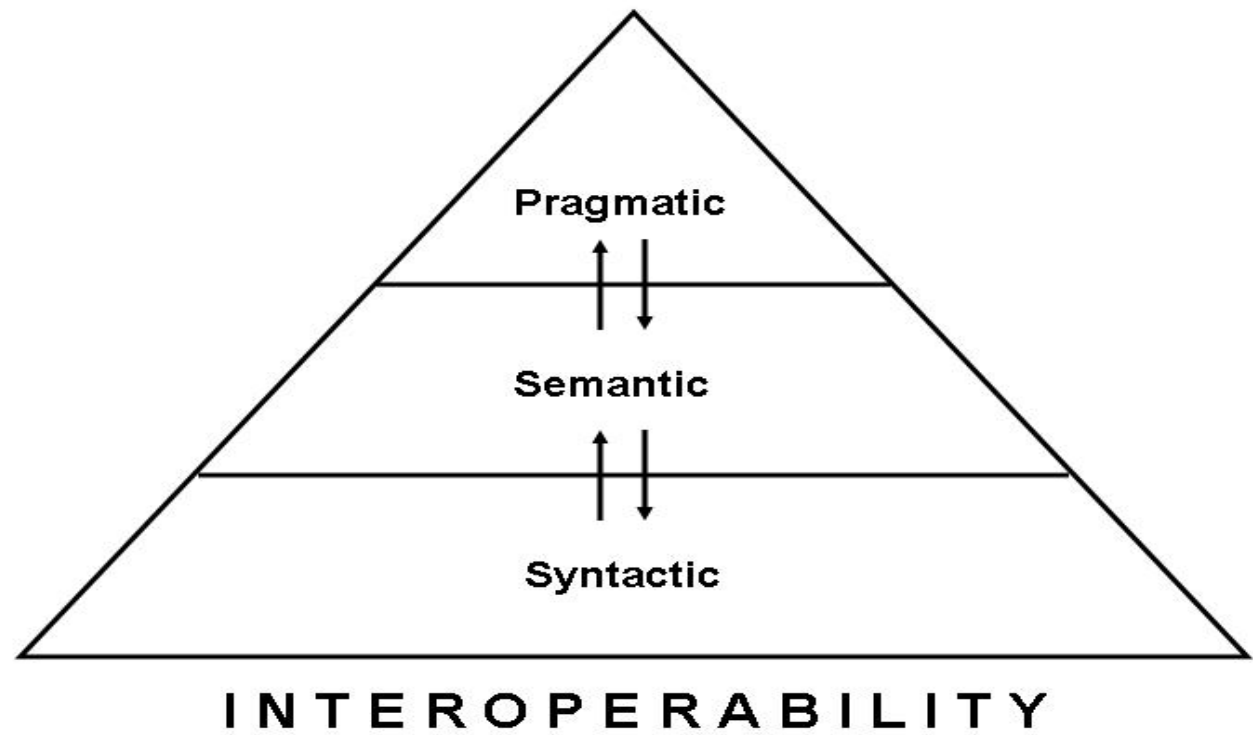
- What is a terminology?
 - Definition of a terminology: a structured set of concepts and terms of a specific subject field in a specific language -> a semantic structure
 - Terminology as an abstract noun denotes the subject field of terminology studies, terminology work, etc. -> terminological semantics (new theories to be implemented in new methodologies (e.g. frame semantics, referential-systemic-functional semantics, how to handle different kinds of indeterminacy, dynamics, diversity, and complexity in domain systems and applications))

Terminology Management as an Instrument of Semantics Management

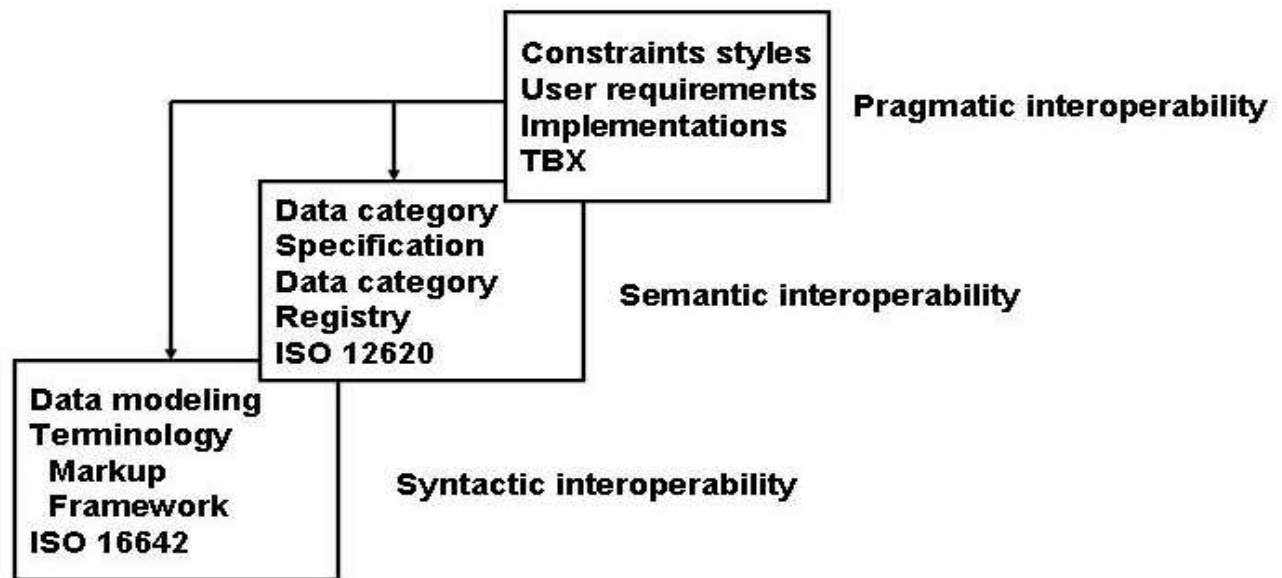
- Terminologies have many different functions in
 - Communication and discourse (mutual understanding)
 - Information (conceptual data models, data storage, logistics)
-> **terminologies are conceptual meta-data**
 - Cognition (concept formation, creative thinking and naming of identified objects in the world)
 - Knowledge (knowledge representation, dissemination, learning, storage, coding, etc.) -> **terminologies are conceptual meta-knowledge of a domain**
 - Professional work in science, technology, industry, business, trade, public and social affairs, culture, etc.
- ALL functions are relevant in semantics management, standardization management and interoperability management

Semantic Interoperability Framework

generic interoperability framework



terminological interoperability



What is Terminology Management?

- A broad concept, covering a wide range of practical activities for manipulating terminological information for specific purposes
- Operationalizes theoretical principles of terminology as methodologies
- A type of Information Management
- Essentially semantics management -> a key to Language Management, Knowledge Management, Content Management

Types of Terminology Management

- **Descriptive** Terminology Management (aiming at documenting terminological diversity, for research purposes or for creating a sound basis for decisions to be taken)
 - Translation-Oriented Terminology Management
(comparative approach including cross-cultural aspects, documenting terminological information structures in source language and target languages)
 - Corpus-driven Terminology Management
(usage-oriented, term extraction from real life discourse, computational terminology paradigm)

Types of Terminology Management

- Managing terminological diversity in social sciences and related disciplines
(e.g. documenting theory-dependent concepts in sociology, contrasting legal terminologies of different countries or legal traditions, socio-terminology)
- Documenting terminological change
(historical approach in order to re-construct the development of terminologies in particular languages or scientific disciplines, in the professions, and social practices)

Types of Terminology Management

- Prescriptive Terminology Management
(a normative approach as part of language planning and technical and scientific standardization, decisions are taken on the basis of existing information sources, terminology collections, etc., aiming at reducing terminological complexity and diversity)
 - Standardizing terminological information
(engineering, natural sciences, medicine, etc.)
 - Standardizing the methods of terminology creation, of terminology management
(term formation, terminography, quality management, process management/workflows)

Approaches (to all types of TM)

- Ad-hoc approach
(aiming at instant problem solving, as part of other processes [translation, technical documentation, technical standardization, etc.])
 - Text-oriented approach
- Systematic approach
(consistent application of work methods, systematic problem solving, interaction of workflows)
 - Domain knowledge-oriented approach

Term Formation in International Terminology Standardization

- The standardization of terminologies has been a prerequisite and a driving force for technical standardization
- Technical standardization is the *raison d'être* of terminology standardization
- At international, European, and national levels, there are thousands of terminology standards (standardized vocabularies) either as autonomous documents or as parts of technical standards in many different subject fields and different languages

Terms, Terminology, Standardization

- In domain communication, a vast number of new technical terms are created every day in hundreds of languages all over the world
- These terms form terminologies as the sets of terms with their specialized meanings (concepts) of a particular domain in a specific language

Terms, Terminology, Standardization

- Terminology standardization has been practiced for thousands of years as a concomitant and necessary feature of the history of science
- Terms may consist of simple words or of complex phrases with specific morpho-syntactic and morpho-semantic features that are unknown in general language but that are specific to certain domains

Term Formation Principles in Terminology Standardization

- Semiotic principles, basically applicable to ‘all’ languages, focusing on the systematic nature of terminologies with their underlying conceptual networks
 - Transparency (vs. opacity)
 - Consistency
 - Appropriateness
 - Conciseness (linguistic economy)
 - Derivability
 - Linguistic Correctness
 - Preference for Native Language

Terminology Standardization

- History of terminology is a history of terminology standardization
- Different degrees of prescriptiveness
- Dynamic vs. static view – benefits and dangers
- Standardization vs. harmonization
- Methodology, research, marketing issues

Areas of standardization

- Terminological principles and methods and Preparation and layout of terminology standards (meta-methodology)
- Vocabulary of terminology (meta-language)
- Terminography (meta-data)
- Terminology documentation (meta-data)
- Coding (meta-data)
- Computational terminology (meta-models)

ISO/TC 37

- Founded in 1936, re-established in 1952
- Title: “Terminology and other language resources”
- Scope: Standardization of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity
- 4 sub-committees (about 4 working groups each) and their scopes:
 - SC 1 Basic principles, knowledge organization, terminology planning
 - SC 2 Methods and workflows, coding, lexicography, translation management, certification
 - SC 3 Computational terminology management, semantic interoperability and data interchange, content management
 - SC 4 Language resource management, annotation, ontologies, etc.

Methodological Standards

- ISO 704 Principles and methods of terminology
- ISO 860 Terminology work – harmonization of concepts and terms
- ISO 1087 Terminology – Vocabulary
- ISO 1087-2 Terminology work – Vocabulary – Part 2: computer applications
- ISO 1951 Lexicographical methods
- ISO 10241 Preparation and layout of international terminology standards
- ISO 12616 Translation-oriented terminography
- ISO 15188 Project management guidelines for terminology standardisation
- ISO/WD 22134 Practical guide for socioterminology
- ISO/NWI TR 22128 Quality assurance guidelines for terminology products
- ISO/NP 23185 Assessment and benchmarking of terminological holdings

Meta-data and Coding Standards

- ISO 639 framework (parts 1-6) Code for the representation of names of languages
- ISO 12199 Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet
- ISO 12615 Bibliographic references for terminology work
- ISO 12620 Computer applications in terminology – Data categories

Interoperability and data modeling Standards

- ISO 16642 Terminology Markup Framework
- ISO 12200 Computer applications in terminology – machine-readable terminology interchange format (MARTIF) – Negotiated interchange
- New: TBX as an ISO Standard

Computational linguistics standards

- ISO/NP 23679-1 Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 1: General principles and methods
- ISO/NP 23679-2 Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 2: Word segmentation for Chinese, Japanese and Korean
- ISO/CD 24610-3 Language resource management – Feature structures – Part 3: Word segmentation for other languages
- ISO/CD 24611 Language resource management – Morpho-syntactic annotation framework
- ISO/CD 24612 Language Resource Management – Linguistic Annotation Framework
- ISO/CD 24613 Language resource management – Lexical markup framework

Standardized Vocabularies

- All domains of technical standardization and harmonization
- Environment, Quality Management, Information Technology, etc.
- Not only in standardization environments (international, regional, national), but also in companies, organizations, networks, etc.
- Increasingly web-based, distributed

LIRICS

Linguistic Infrastructure for Interoperable Resources and Systems

GOALS:

- LIRICS provides a common standards framework for language engineering by translating requirements from European language industry into ISO standards on the basis of ongoing R&D work
- LIRICS provides input,
 - on the basis of the cooperation and interaction between research consortia and industry groups,
 - to ongoing standards work in ISO/TC 37,
 - focusing on lexicons, morpho-syntax, syntax and semantic content.
 - accompanied by a set of test suites in nine European languages to facilitate their implementation and an open source implementation platform
 - allowing common-format, multi-lingual language processing compatible with legacy systems and tools

Primary resources

Texts, spoken data,
multimedia information
[TEI, MPEG7, TMX,
XHTML, etc.]

Access protocols

[Corba, SOAP]

Knowledge structures

Hierarchies of types
Relations between concepts
SKOS [Topic Maps,
RDF/RDFS/OWL]

LIRICS domain of impact

LIRICS scope

NLP structures

Linguistic annotations
Tokenisation
Morpho-Syntactic Tagging
Chunks (e.g. Named Entities, etc.)
Deep syntactic structures
Co-references etc.
[Eagles, ISLE, Multext/Multext-East
CES, MATE, Whiteboard]

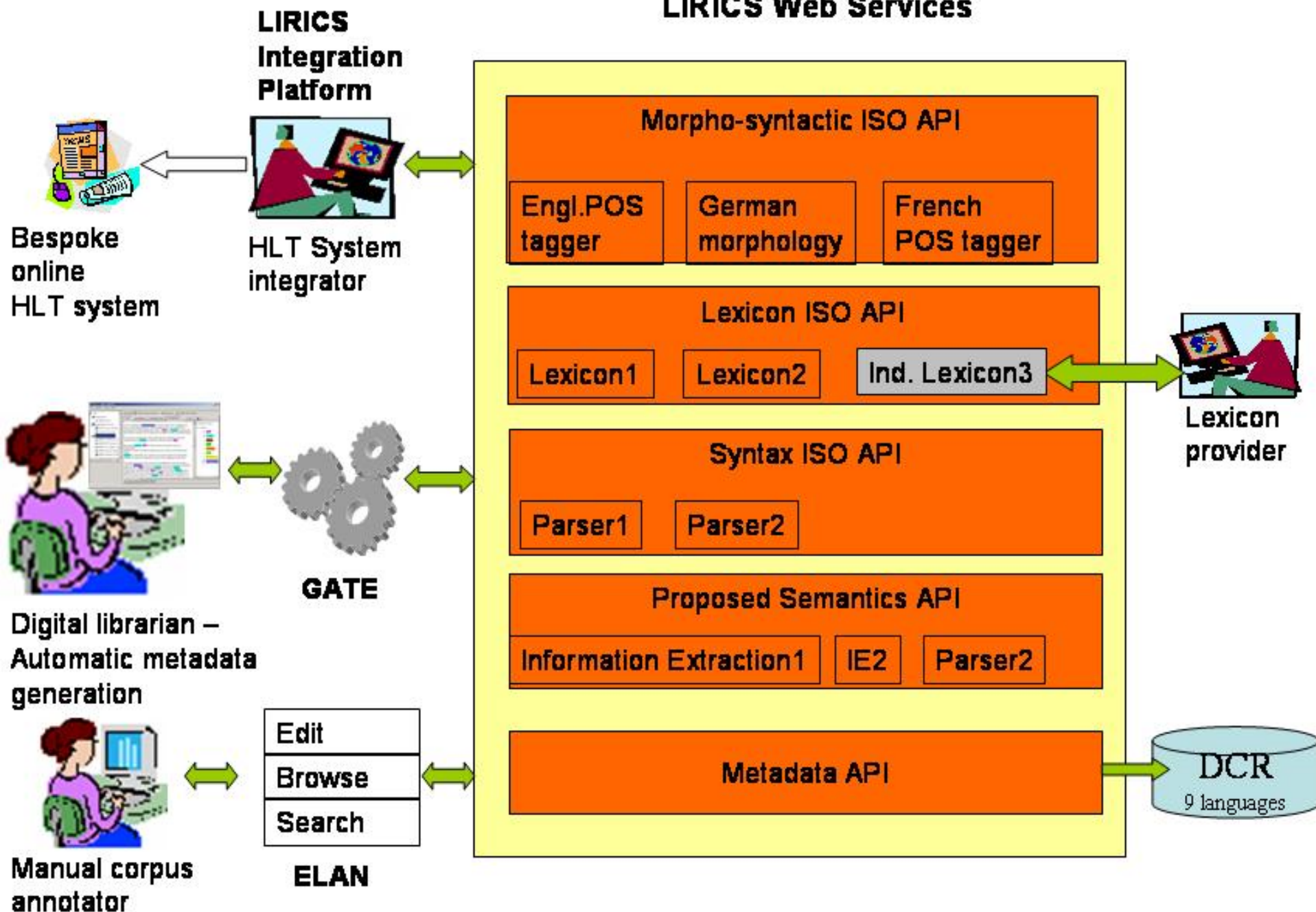
Meta-data

[Dublin core, TEI,
OLAC, IMDI, MPEG7]

Lexical structures

Terminologies
Morphological lexica
Syntactic lexica
Transfer lexica
[ISO 16642, TBX, OLIF,
Genelex/Simple/ISLE]

LIRICS Web Services



Language Resource Management Standardization

- Standardization is needed for language resources (mono- and multilingual), e.g. speech data, written (full) text corpora, lexical (general language) corpora and their processing methods
- Relevant research areas are computational linguistics and computational lexicography, language engineering, etc., which have provided industrial best practices to be turned into official standards
- This process will contribute to the further development of the language industries at large
- As is the case with terminologies, language resources in general are often multilingual, multimedia and multimodal

ISO/TC 37/SC 4

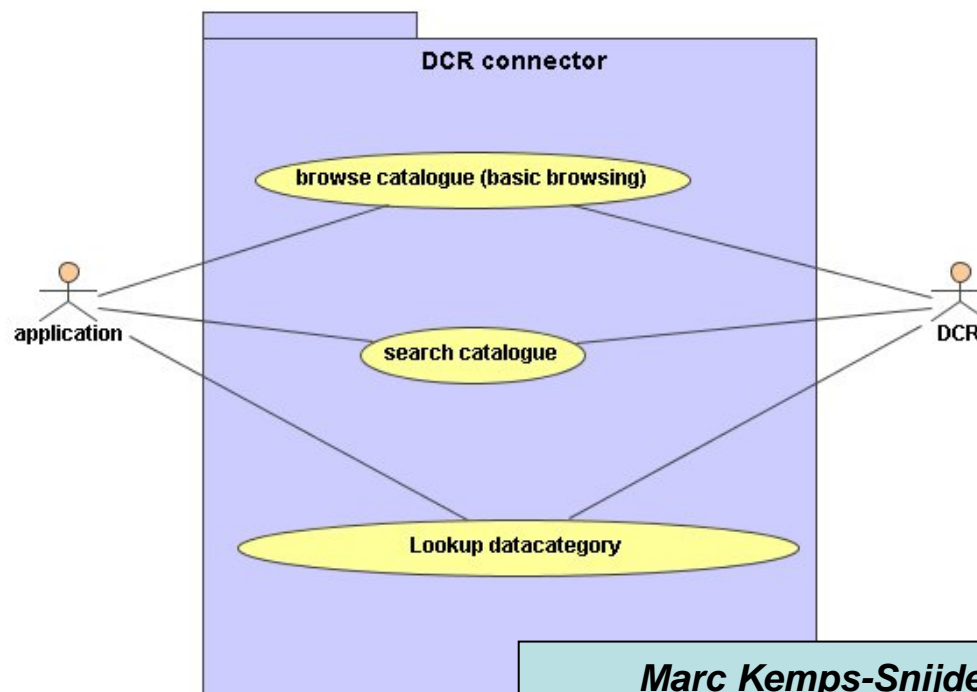
- Title: Language resource management
- Scope: Standardization of specifications for computer-assisted language resource management
- linguistic infrastructures are being established or re-enforced as part of the rapidly evolving information and communication society;
- professional activities involving language resource sharing and standardization are increasing in diverse areas:
 - governmental or non-governmental organizations, public or private institutions, educational institutions, commercial enterprises, etc.,
 - both, globalization and localization necessitate multilingual communication;
- there is an increasing need for new standardization as well as urgent recognition of existing de facto standards and their transformation into International Standards

Interface to DCR (ISO 12620) service

- Purpose is to achieve a higher degree of interoperability between linguistic resources by providing an interface for external applications.

- Use cases

- Browse DCR
- Search DCR
- Lookup datacategory



Browsing the DCR

- function List getProfiles()
 - Returns a list of all profiles (e.g. metamodel, terminology, morphosyntax etc.)
- function List getDataCategories(a_profile,a _registrationStatus)
 - Returns a list of all datacategories associated with the specified profile having the specified registration status
- function DataCategory getDataCategory(URID)
 - Returns a datacategory identified by the specified ID

Browsing the DCR

DataCategory details homophone - Microsoft Internet Explorer provided by MPI Nijmegen

Data category information

homophone

(MorphoSyntax)

Administration Identification

Identifier: homophone
Version: 0.0.0
Registration authority: Private
Registration status: candidate
Administration status: Private
Origin: ?

Creation date: 2004-07-09
 ?

Last change date: 0000-00-00
 ?

Description

Profile: MorphoSyntax

Definition	Source	Note
Un mot qui se prononce de la même manière qu'un autre mot mais qui s'écrit différemment. Gil Francopoulo		

Language section english

Name section		
Name	Status	Note
homophone	standardizedName	

Language section french

Name section		
Name	Status	Note
homophone	standardizedName	

Marc Kemps-Snijders
 MPI for Psycholinguistics
Adam Funk
 University of Sheffield

DCR web service implementation

- DCR service is implemented as REST web service
 - E.g. http://syntax.inist.fr/mod_webservice/call.php?fct=getProfiles
- Response is provided in RelaxNG format.
 - E.g.

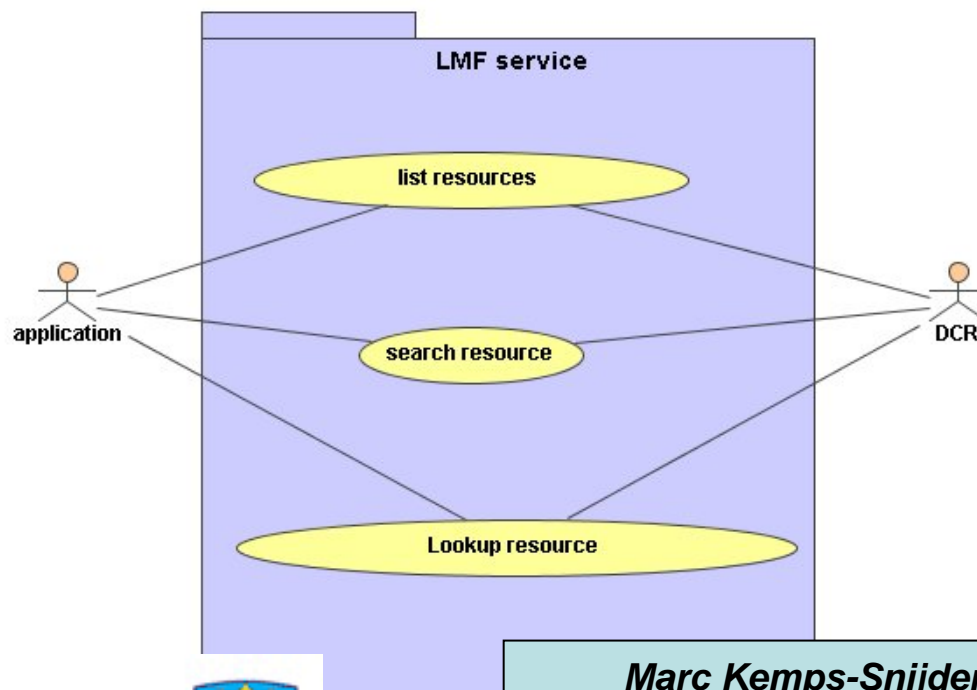
```
<element name="STRUCT">
  <attribute name="type">
    <value>ListofProfiles</value>
  </attribute>
  <zeroOrMore>
    <element name="feat">
      <attribute name="type">
        <value>profile</value>
      </attribute>
      <text/>
    </element>
  </zeroOrMore>
</element>
```

Interface to LMF service

- Purpose is to allow access to content from lexical resources by external applications

- Use cases

- List resources
- Search resource
- Lookup resource



LMF web service implementation

- LMF service is implemented as SOAPweb service
- WSDL available for client generation
- Lexicon documentation is generated at namespace identifier.
- Response format is well defined
 - E.g

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">  
<xs:element name="resources">  
  <xs:complexType>  
    <xs:sequence>  
      <xs:element maxOccurs="unbounded" ref="lexicon"/>  
    </xs:sequence>  
  </xs:complexType>  
</xs:element>  
<xs:element name="lexicon">  
  <xs:complexType>  
    .....
```
 - Schema information is returned in XML Schema (RelaxNG is also possible)
 - Resource information and search results validate against resource (sub) schema.

Gate LMF plugin

- Visual Resource similar to the WordNet plug-in
- WSDL back-end according to the LMF standard
- Standards
 - ISO/TC 37/SC 4 LMF
 - Unicode
 - ISO 12620 DCR
 - GATE (one of the *de facto* standards in NLP)

- LexicalEntry
 - LemmatisedForm
 - InflectedForm
 - RepresentationFrame (alternative orthography)
- EntryRelation
 - Headword <-> RelatedForm
- Definition, Translation
- SenseRelation

Standards infrastructures

ISO 16642 TMF	ISO 24611 MAF	ISO 24612 LAF	ISO 24613 LMF	ISO 24615 SynAF	ISO nnnnn SemAF
Terminological, Morphsyntactic, Linguistic, Lexical, Syntactic and Semantic Data Categories (Metadata)					
ISO 12620: Data Categories					
ISO 639-1-6 (x4) Language Codes					
ISO 639-4: Language Code metamodel					
ISO 11179: Metadata Registries					



Messages file:/H:/Documents/Gate Datastore/ REVEAL FirstExpertTranscripts Corpus Pipeline_00023

Annotation Sets Annotations Co-reference Editor Text

Type	Set	Start	End	Features
DiscoveredTerm		101	101	{Term=, ValidLinguistic=false, ValidStatistical=true}
DiscoveredTerm		106	119	{Term=shopping mall, ValidLinguistic=true, ValidStatistical=false}
DiscoveredTerm		144	155	{Term=individuals, ValidLinguistic=true, ValidStatistical=false}

2029 Annotations (0 selected)

08.18
 Man in blue t-shirt stretches in the centre of the floor facing the camera. Exits. Second man, in black trousers and grey top, walks across shopping centre to the far side, sits on the floor in front of something for a few seconds. Gets up and exits (inaudible) screen.

09.31
 Man in blue t-shirt holds up document to the camera again. Two males, one in red, one in a black top with yellow stripes meet, brief scuffle in the centre of the shopping centre. Both run off in the same direction but through different exits.

10.02
 It's a shop front, first storey shopping mall. Person of unknown gender leaves the shop and a man wearing a white shirt enters.

10.36
 I think it's a male leaving the shop, long hair, exits to the left.

10.55
 Street scene. Broadway Street Church, apparently. White van passes. Four parked cars, one white van parked. Female loitering on far right by a hedge.

11.36
 Could be male loitering, I don't know.

- DiscoveredTerm
- Keyword
- LinguisticTerm
- Sentence
- SpaceToken
- Split
- StatisticalTerm
- Token

New

Document Editor Initialisation Parameters

GATE annotation framework

Conclusions and outlook

- Industry requirements -> specifications for designing standards and testing them for real-world use (close interaction in the workflow)
- Feedback loops/co-operation schemes/liaison strategies/workshops
- Critically reviewing and revising existing standards to live up to changing and new expectations and requirements

- The full potential of TC 37 standards can only be exploited when they are linked to other standards, thereby contributing to building **standards ecosystems, e.g.**
 - The ISO 11179 Framework has become an over-arching meta-standard for all TC 37 standards and many other standards
 - TC 37 standards have become meta-standards for 11179 standards and many other standards (e.g. ISO 14000)
- All TC 37 standards will have to create a digital presence -> web-based, distributed digital infrastructures (such as ISO 12620, ISO 639, etc.)