

# A new Generation of Language Industry Standards – the LIRICS Project

2007-05-10

IAB Meeting

AFNOR

Gerhard Budin  
University of Vienna

## 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

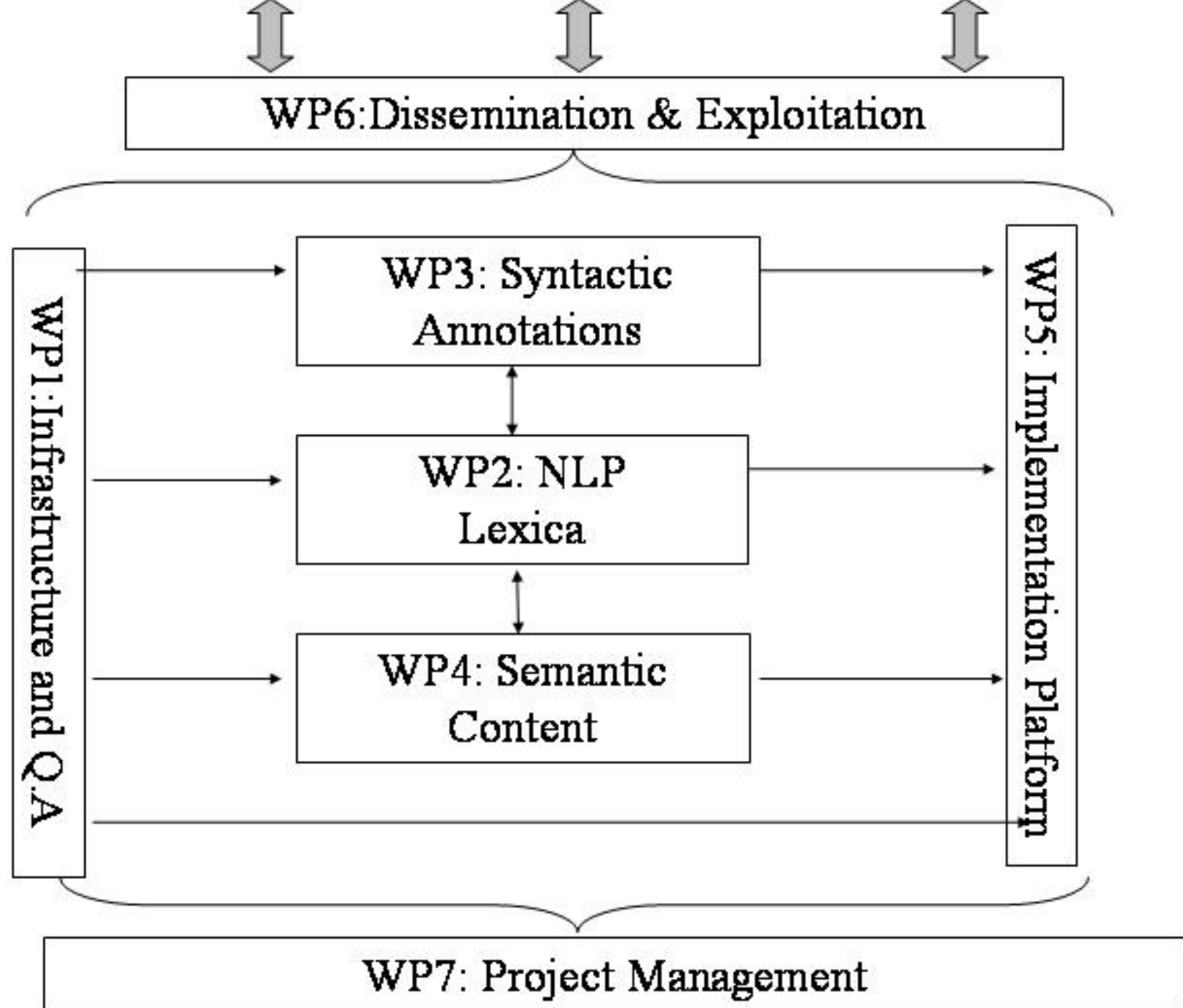
Linguistic annotations  
Tokenisation  
Morpho-Syntactic Tagging  
Chunks (e.g. Named Entities, etc.)  
Deep syntactic structures  
Co-references etc.  
[Eagles, ISLE, Multext/Multext-East  
CES, MATE, Whiteboard]

### Meta-data

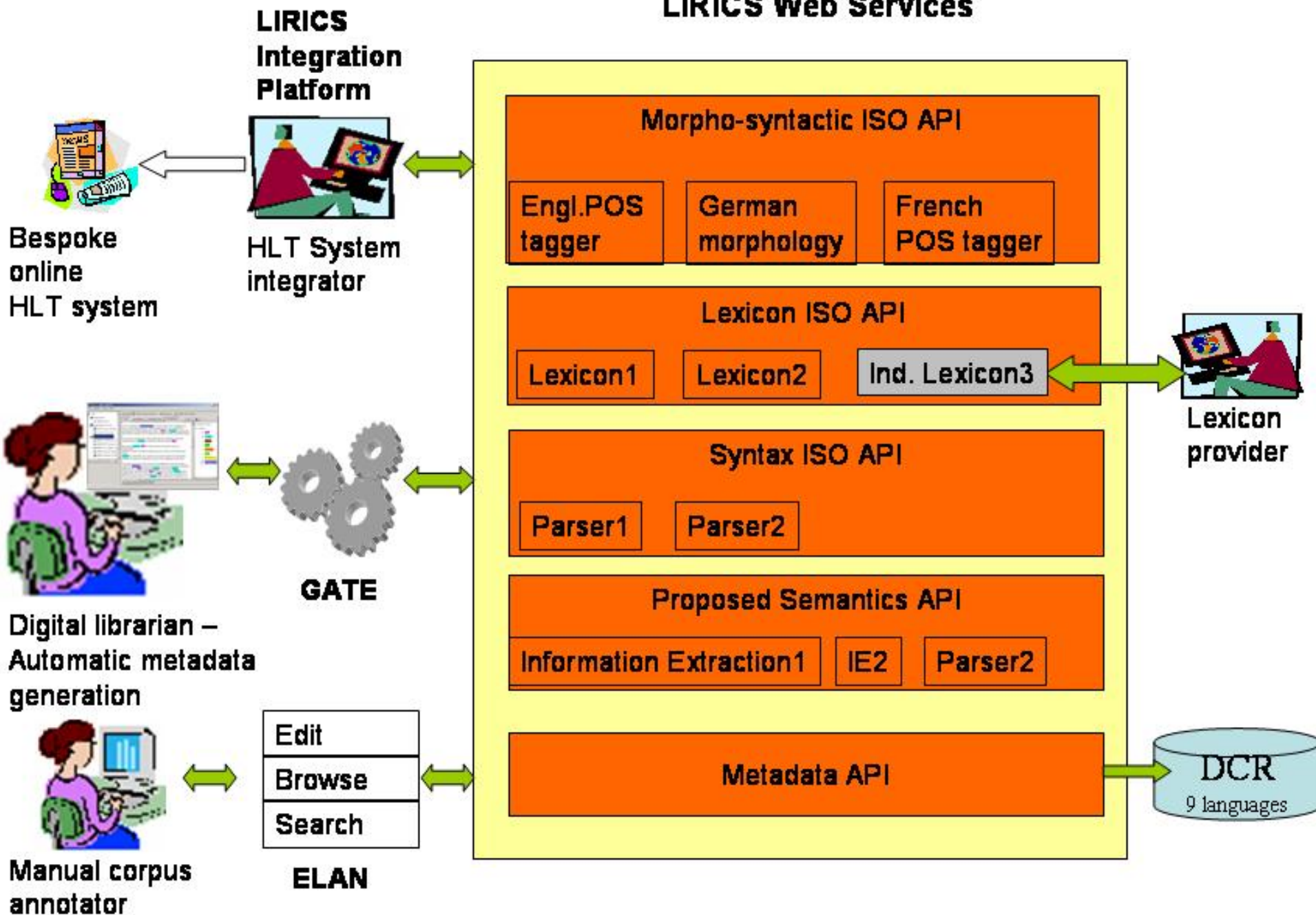
[Dublin core, TEI,  
OLAC, IMDI, MPEG7]

### Lexical structures

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



# LIRICS Web Services



## Creating ISO standards

1. Viable idea, existing documents (including de-facto industry standards) representing real needs and requirements from society (industry/trade, consumers, research, social and cultural institutions, etc.)
2. National standards committees (AFNOR, BSI, DIN, AENOR, ON, etc.) or international committees (ISO) present a New Work Item Proposal (NWIP) for vote (certain requirements to be fulfilled)
3. Assignment of the NWI to a working group of a (sub-)committee, NWI to be edited by project editor in cooperation with a project team within a working group (experts to be nominated by national member committees, plus liaison representatives)
4. Presentation of WD (Working Draft) for vote to become a CD (Committee Draft), receiving comments to be resolved for presenting the CD for vote to become a DIS (Draft International Standard), receiving comments to be resolved for presenting the FDIS (Final Draft International Standard) to become an IS (International Standard)
5. Standards to be reviewed and updated at regular intervals
6. Fast-track procedure, Vienna Agreement (CEN-ISO)

# ISO/TC 37

## Terminology and other language and content resources

- Founded in 1936/re-established in 1951
- 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
  - SC 1 Principles and Methods (chair: L.-J. Rousseau, Secr. Sweden)
  - SC 2 Terminography and Lexicography (chair: G. Budin, Secr. Canada)
  - SC 3 Computer applications (chair: B. Nistrup Madsen, Secr. Germany)
  - SC 4 Language Resource Management (chair: L. Romary, Secr. Korea)
- Each SC has several working groups which run at least one project
- Based on practical needs horizontal cooperation and coordination is to be guaranteed by SC chairs

## 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 1

The following standards are under the direct responsibility of ISO/TC 37/SC 1:

- ISO 704:2000 Terminology work – Principles and methods
- ISO 860:1996 Terminology work – Harmonization of concepts and terms
- ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application

The following standards are under preparation:

- ISO/CD 704 Terminology work – Principles and methods
- ISO/CD 860 Terminology work – Harmonization of concepts and terms
- ISO/PWI 1087-1 Terminology work – Vocabulary – Part 1: Theory and application
- ISO/WD 22134 Practical guide for socioterminology

## ISO/TC 37/SC 2

- Title: Terminography and lexicography
- Scope: Standardization of terminological and lexicographical working methods, procedures, coding systems, workflows, and cultural diversity management, as well as related certification schemes

## ISO/TC 37/SC 2 (2)

The following standards are under the direct responsibility of ISO/TC 37/SC 2:

- ISO 639-1:2002 Codes for the representation of names of languages – Part 1: Alpha-2 code
- ISO 639-2:1998 Codes for the representation of names of languages – Part 2: Alpha-3 code
- ISO 1951:1997 Lexicographical symbols and typographical conventions for use in terminography
- ISO 10241:1992 International terminology standards -- Preparation and layout
- ISO 12199:2000 Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet
- ISO 12616:2002 Translation-oriented terminography
- ISO 15188:2001 Project management guidelines for terminology standardization

## ISO/TC 37/SC 2 (3)

The following standards are under preparation:

- ISO/DIS 639-3 Codes for the representation of names of languages Part 3: Alpha-3 code for comprehensive coverage of languages
- ISO/CD 639-4 Codes for the representation of names of languages Part 4: Implementation guidelines and general principles for language coding
- ISO/CD 639-5 Codes for the representation of names of languages Part 5: Alpha-3 code for language families and groups
- ISO/WD 639-6 Codes for the representation of names of languages Part 6: Extension coding for language variation
- ISO/FDIS 1951 Presentation/representation of entries in dictionaries
- ISO/CD 10241-1 Terminological entries in standards – Part 1: General requirements
- ISO/CD 10241-2 Terminological entries in standards
- ISO 12615 Bibliographic references and source identifiers for terminology
- ISO/NWI TR 22128 Quality assurance guidelines for terminology products
- ISO/NP 23185 Assessment and benchmarking of terminological holdings

## ISO/TC 37/SC 3 (1)

- title: Terminology management systems and content interoperability
- scope: Standardization of principles and requirements for semantic interoperability, terminology and content management systems, and knowledge ordering tools

## ISO/TC 37/SC 3 (2)

The following standards are under the direct responsibility of ISO/TC 37/SC 3:

- ISO 1087-2:2000 Terminology work – Vocabulary – Part 2: Computer applications
- [ISO 6156:1987](#)  
(withdrawn) Magnetic tape exchange format for terminological/ lexicographical records
- ISO 12200:1999 Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange
- ISO 12620:1999 Computer applications in terminology – Data categories
- ISO 16642:2003 Computer applications in terminology – Terminological markup framework

## ISO/TC 37/SC 3 (3)

The following standards are under preparation:

- ISO/NWI TR 12618 Computational aids in terminology – Design, implementation and use of terminology management systems
- ISO/CD 12620 Computer applications in terminology – Data categories

## ISO/TC 37/SC 4 (1)

- 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

## ISO/TC 37/SC 4 (2)

The following standards are under preparation:

- ISO/NWI 21829 Terminology for language resources
- 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/WD 24611 Language resource management – Morpho-syntactic annotation framework
- ISO/WD 24612 Language Resource Management – Linguistic Annotation Framework
- ISO/WD 24613 Language resource management – Lexical markup framework

## Conclusions

- Industry requirements
- -> specifications for designing the standards
- Scoping: language technology standards, terminology and language resource standards, etc.
- Feedback loops/co-operation schemes/liaison strategies/workshops
- Critically reviewing and revising existing standards to live up to changing and new expectations and requirements