

*Language Resource Management*  
*Descriptors and Mechanisms for Language Resources*

File ID wg1N08.doc (544 ko)  
wg1N08.pdf (90 ko) wg1N08.ppt (184 ko) wg1N08.zip (218 ko)

Title :	Meta-data for language resources Slides (slides version)
Editor(s) :	Laurent Romary
Source :	WG 1
Project number :	N/A
Status :	Contribution – Slide show presented during the Vienna meeting Austria 2002-08-19/23 (done by L. Romary).
Date :	2002-08-22
Agenda / Action :	For consideration, discussion and comment
References :	Based on SC4 WG1 N02 contribution

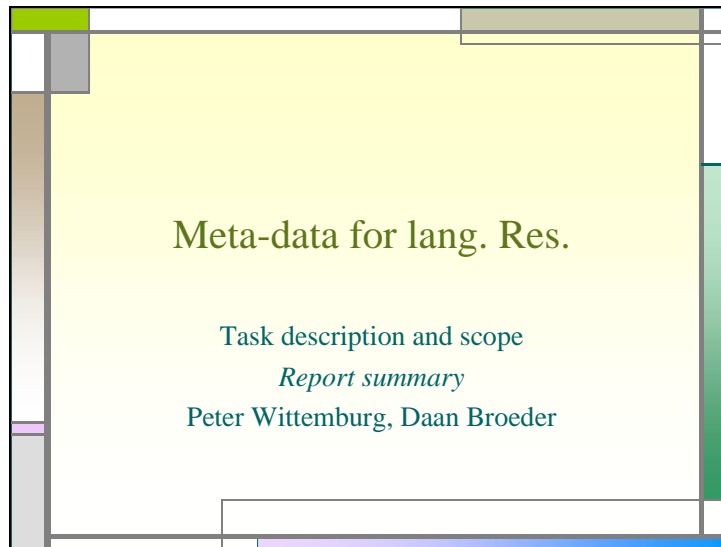
Mr. Key-Sun Choi - SC4 Secretary – KORTERM - 373-1 Kusong-Dong Yusong-gu - Taejon 305- 701 - Korea  
82 42 869 35 25 – fax: 82 42 867 35 65 - [kschoi@cs.kaist.ac.kr](mailto:kschoi@cs.kaist.ac.kr) – <http://korterm.kaist.ac.kr>

---

Any question about SC4/WG1 documents registration, maintenance and distribution must be forwarded to Afnor  
France preferably to: [sylvie.arbouy@afnor.fr](mailto:sylvie.arbouy@afnor.fr)

---

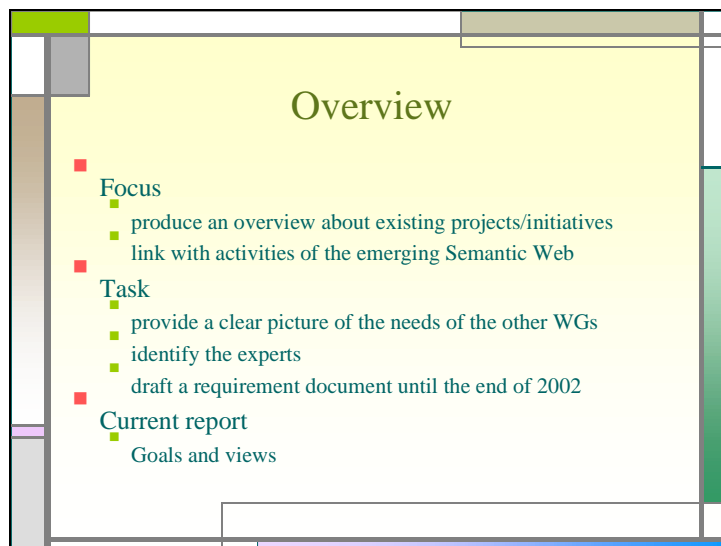
Slide 1



Meta-data for lang. Res.

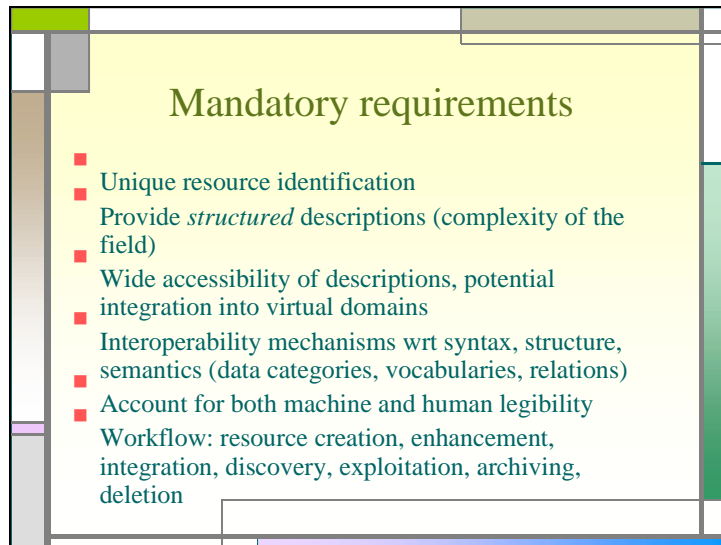
Task description and scope  
*Report summary*  
Peter Wittemburg, Daan Broeder

Slide 2



Overview

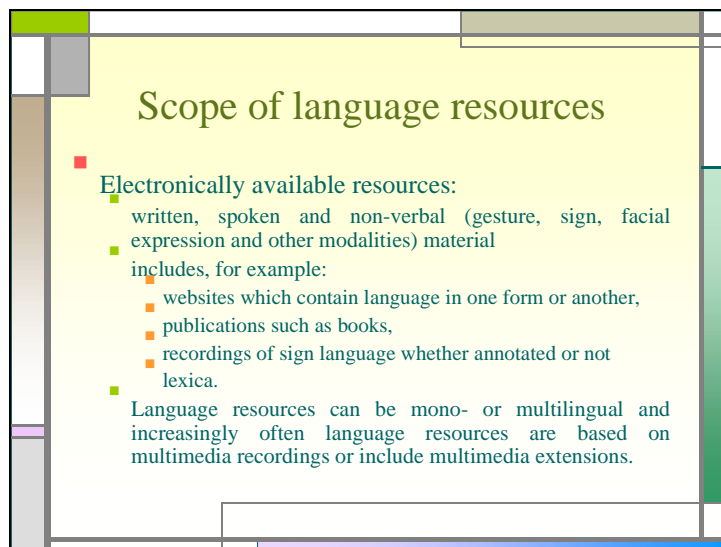
- Focus
  - produce an overview about existing projects/initiatives
  - link with activities of the emerging Semantic Web
- Task
  - provide a clear picture of the needs of the other WGs
  - identify the experts
  - draft a requirement document until the end of 2002
- Current report
  - Goals and views



A slide with a yellow background and a decorative border. The title "Mandatory requirements" is centered at the top in a green serif font. Below the title is a list of six items, each preceded by a red square bullet point. The text is in a dark green serif font.

### Mandatory requirements

- Unique resource identification
- Provide *structured* descriptions (complexity of the field)
- Wide accessibility of descriptions, potential integration into virtual domains
- Interoperability mechanisms wrt syntax, structure, semantics (data categories, vocabularies, relations)
- Account for both machine and human legibility
- Workflow: resource creation, enhancement, integration, discovery, exploitation, archiving, deletion



A slide with a yellow background and a decorative border. The title "Scope of language resources" is centered at the top in a green serif font. Below the title is a list of items. The first item is "Electronically available resources:" followed by a list of sub-items. The text is in a dark green serif font.

### Scope of language resources

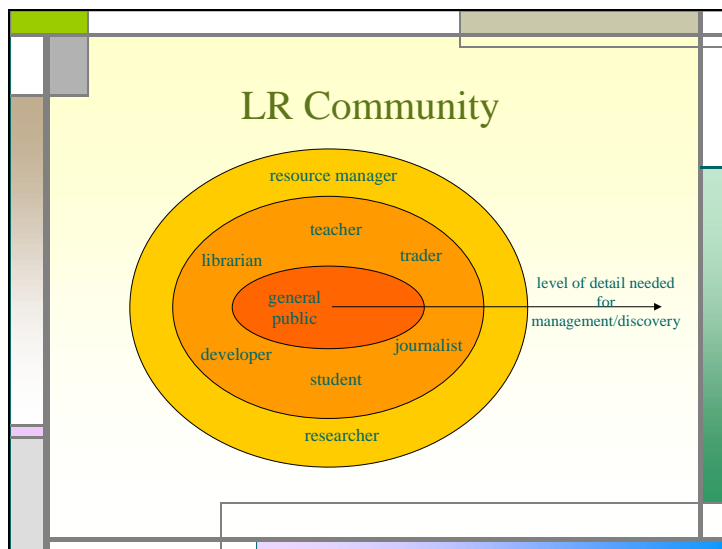
- Electronically available resources:
  - written, spoken and non-verbal (gesture, sign, facial expression and other modalities) material
  - includes, for example:
    - websites which contain language in one form or another,
    - publications such as books,
    - recordings of sign language whether annotated or not
  - lexica.
- Language resources can be mono- or multilingual and increasingly often language resources are based on multimedia recordings or include multimedia extensions.

Slide 5

## Scope of language resources

- **Derived types:**
  - linguistic resources that contain metadata about other language resources in the wider sense
  - Lexica, grammar notes and many other types of language resources contain abstract linguistic material and refer to more basic types of resources such as annotated recordings.
- **Restrictions:**
  - **Orientation towards the study of language**
    - Many other possible views on linguistically based documents (car manual, movie script, etc.)
  - Metadata descriptions describing language resources with a set of typical categories are not meant to be language resources themselves

Slide 6

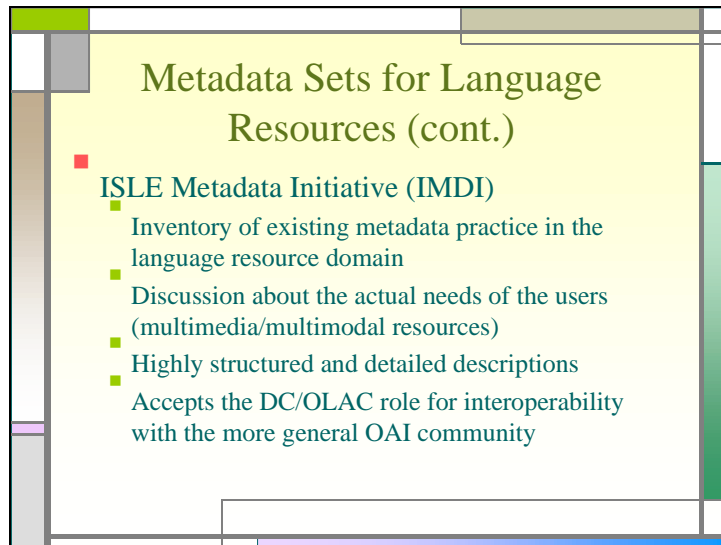


## Relevant initiatives

- Header information/legacy meta-data
  - E.g. CHAT header of Childes project, TEI header
  - Mainly dedicated to textual information
- Dublin Core Metadata element set
  - Influenced by the librarian community
  - Links with ISO 11179 (specification) and XML/RDF (implementation)

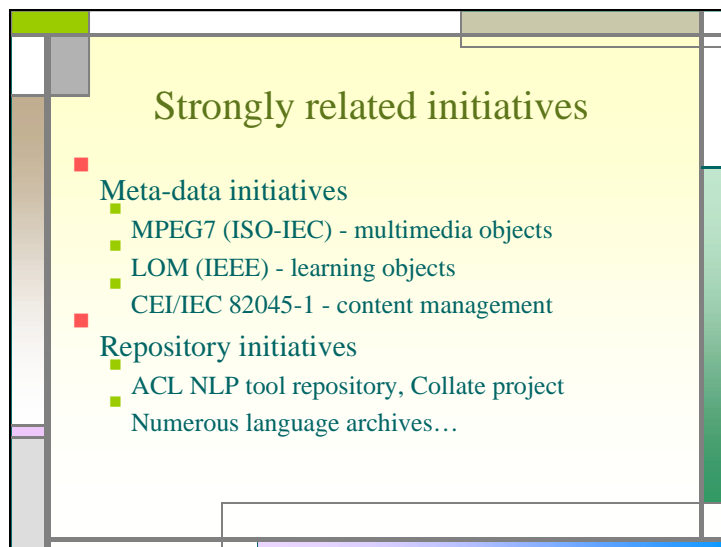
## Metadata Sets for Language Resources

- OLAC (Open Language Archives Community)
  - Motivators from LDC and SIL
  - Addition of a number of sub-elements to DCMES
    - E.g.: object language, linguistic data type
  - Definition of a number of refinements of the element semantics.
  - Flat structure, wide coverage of language resources types



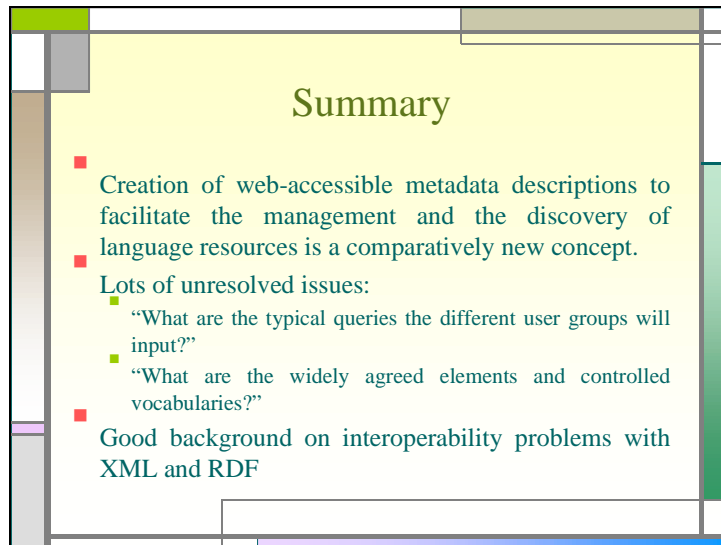
Metadata Sets for Language Resources (cont.)

- ISLE Metadata Initiative (IMDI)
  - Inventory of existing metadata practice in the language resource domain
  - Discussion about the actual needs of the users (multimedia/multimodal resources)
  - Highly structured and detailed descriptions
  - Accepts the DC/OLAC role for interoperability with the more general OAI community



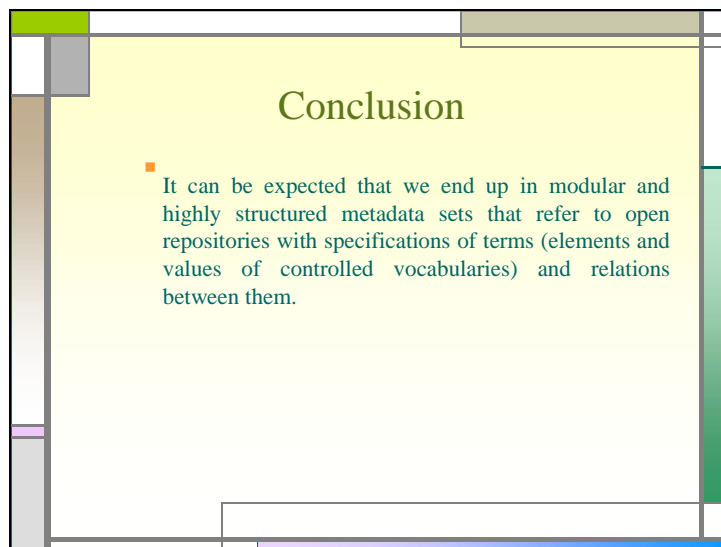
Strongly related initiatives

- Meta-data initiatives
  - MPEG7 (ISO-IEC) - multimedia objects
  - LOM (IEEE) - learning objects
  - CEI/IEC 82045-1 - content management
- Repository initiatives
  - ACL NLP tool repository, Collate project
  - Numerous language archives...



Summary

- Creation of web-accessible metadata descriptions to facilitate the management and the discovery of language resources is a comparatively new concept.
- Lots of unresolved issues:
  - “What are the typical queries the different user groups will input?”
  - “What are the widely agreed elements and controlled vocabularies?”
- Good background on interoperability problems with XML and RDF



Conclusion

- It can be expected that we end up in modular and highly structured metadata sets that refer to open repositories with specifications of terms (elements and values of controlled vocabularies) and relations between them.

## Actions

- define simple schemas that are used to define elements and vocabularies
- enforce agreements on a number of controlled vocabularies
- take care that all elements and vocabularies used in IMDI and OLAC are well-defined in accordance with the schemas
- take care that these elements are available via open repositories
- work out the Semantic Web scenario.