



ISO TC 37/SC4/WG1 N02  
2002-08-09

*Language Resource Management*  
*Descriptors and Mechanisms for Language Resources*  
File wg1N02.doc (104 ko)  
wg1N02.pdf (64,4 ko)

Title : WI-3 Task Description and Scope

Editors : Peter Wittenburg, Daan Broeder

Source : WG 1

Project number : Work item 3 (*temporary reference*)

Status : SC4 - Internal draft version - Revision 1.0

Date : 2002-07-15

Agenda : For consideration, discussion and comments during the Vienna WG1 meeting.

Remarks : This version does not include references. The reader is supposed to know the initiatives and frameworks cited. As soon as possible references will be added.

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)

---

# WI-3 Task Description and Scope

- SC4-internal draft version -

## 1. Task

ISO TC37/SC4 is dedicated to improving the management of language resources in a distributed and interlinked scenario on the World-Wide-Web.

The tasks were recently discussed and defined during a Constituent Meeting of TC37/SC4. The official report from the TC37/SC4 meeting specifies

- as focus:
  - produce an overview about existing projects/initiatives and monitor its usage
  - link with activities of the emerging Semantic Web
- as tasks:
  - provide a clear picture of the needs of the other WGs
  - identify the experts
  - draft a requirement document until the end of 2002

The TC37/SC4 resolution document adds another point explicitly: create a basic paper on goals and views. To create the requirement document a task force was installed.

To be able to achieve the goals WI-2 has to

- determine the scope of language resources
- determine the needs of the community and in the realm of TC37/SC4 of the other working groups
- determine the existing initiatives relevant to the language resource domain,
- develop a scenario how metadata will be used in the Semantic Web
- determine the set of descriptors and their vocabularies useful to describe language resources
- define all relevant terminological units (concepts and terms in major languages) and their relations
- define suitable frameworks for the definitions

A number of mandatory requirements need to be fulfilled to establish a manageable domain of language resources:

- All resources have a unique identifier conforming to the web standards (URI).
- Due to the inherent complexity and large spectrum of different types of language resources, metadata descriptions are created which describe their major characteristics to facilitate management. It is assumed that only structured descriptions will provide the necessary precision for managing language resources efficiently.
- Although the resources themselves will not always be openly accessible due to commercial, ethical or legal reasons, the metadata descriptions have to be accessible and must have the potential to be integrated into virtual management domains.
- All items in such metadata domains must adhere to interoperability mechanisms, i.e. syntax, structure, semantics (data categories, vocabularies, relations) have to be described and stored so that humans and programs can use them.
- The description level should not neglect the need for quick inspection by human readers by providing the possibility of entering prose text.
- The term “management” covers the complete workflow cycle, i.e. WI-2 has to consider the processes of resource creation, enhancement, integration, discovery, exploitation, archiving and deletion.
- The Web is international therefore multilinguality is an inherent characteristic and requirement of the language resource domain.

## 2. Scope of Language Resources

Only electronically available resources are included. Further, the term “Language Resource” covers at first instance all resources that contain written, spoken and non-verbal (gesture, sign, facial expression and other modalities) material. This definition includes, for example, all websites which contain language in one form or another, publications such as books, recordings of sign language whether annotated or not and lexica. Language resources can be mono- or multilingual and increasingly often language resources are based on multimedia recordings or include multimedia extensions.

There are many types of 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. Also these derived data types are language resources.

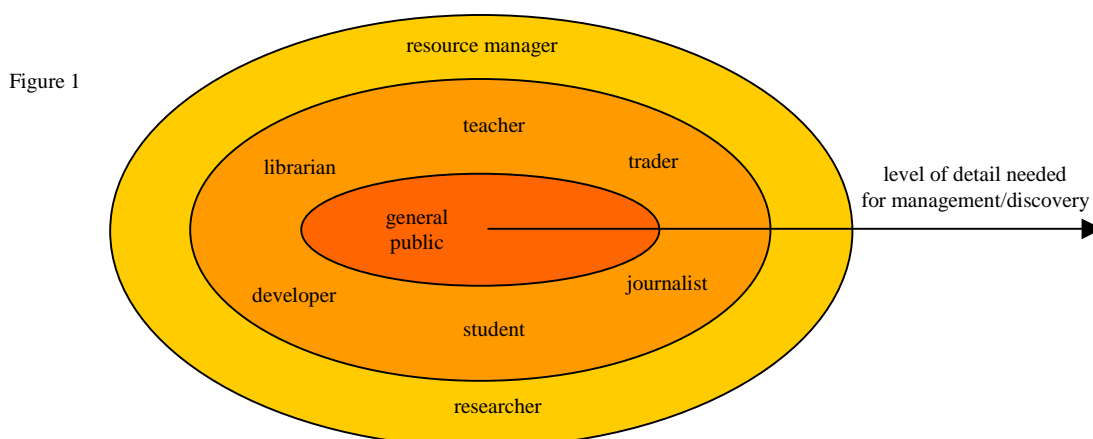
Important for the discussions in this note is the view people have on language resources. Technical documentation of cars, material that is part of learning objects and annotated film movies are all language resources in the above-mentioned broad sense. Nevertheless, completely different communities are involved to search for them. In the case of technical documentation engineers may need certain descriptors to easily retrieve a relevant document. A linguist could look at the same document from a different perspective - a research perspective for example, i.e. he will need other type of descriptors than the engineer. We can assume that especially the description of the content will differ, but until now there is not enough knowledge to make final conclusions.

Fact is that when comparing different sets such as Dublin Core and CEI/IEC 82045-1 from the field of document management for example, that the differences are relatively large although both speak about documents with texts covered in them. Therefore, we need to restrict the scope of "Language Resources" essentially to those directed towards the study of language. Resources not providing this are not part of the domain. Metadata descriptions have to support this view.

The metadata descriptions describing language resources with a set of typical categories are not meant to be language resources themselves in the context of this note. They are solely used for resource management and discovery purposes. This, however, does not exclude that they will be viewed as LRs within the context of other work.

### 3. LR Community

Language resources in the above sense are of interest to many different groups and they fulfill different functions for these groups. Given this variety, it is impossible to define a complete inventory of usage scenarios for LR. We can only identify a few key communities with typical usage interests. On the one hand there is the general public, which is interested in general information on many subjects. On the other hand there are resource traders, researchers or language engineers interested in selling a particular type of language resource, deriving a new grammar or calculating the parameters of statistical recognition algorithms. While the former may be content with general information, the latter will require finer details. Consequently, we can define levels of abstractions as indicated in the following figure. The second level includes many different groups of people, since their usage profile is not at all clear at this moment.



### 4. Relevant Initiatives

There are a number of standards and initiatives that are relevant to the task of WI-2 that will be briefly mentioned in this chapter.

#### Header Information/Legacy Metadata

In recent years, many language resource projects and tools supported typical metadata information describing some main characteristics of the resources. An example is the *CHAT* header of the CHILDES project. Much experience has been assembled when the well-known *TEI* header standard was worked out. Most of the initiatives such as

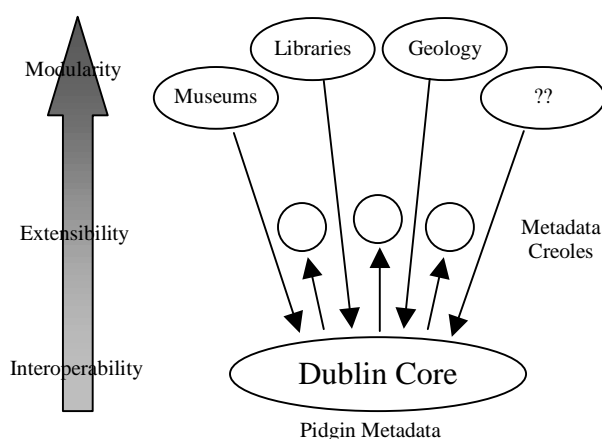
CHILDES dealt with language corpora covering only texts. Multimedia recordings as integral part or as extensions were not used when these formats were defined. Other data types such as lexica were insufficiently described. Also TEI was defined before integral media components became normal. For linguistic data types such as lexica or field notes no particular suggestions were made. Usually the header information was not used for automatic discovery, but for data management purposes. Each project defined its own individual set, since there was no plan to integrate those. Nevertheless, the experiences from these initiatives and efforts have to be considered when deriving metadata standards for language resources. An overview of such initiatives and efforts can be found in the overview document that was the basis for the IMDI metadata set.

### Dublin Core Metadata Element Set

A large number of metadata initiatives were started in the last decade to describe resources of various domains. The **Dublin Core Metadata Element Set (DCMES)** is the most important one, as it claims to be useful to describe all type of web-resources using 15 elements so that they can be easily discovered. DCMES was primarily designed by the librarian community that had a lot of experience with categorical descriptions of all sorts of publications. Furthermore, the *PICs* metadata initiative influenced the DCMES development. There are many domain-driven extensions of DCMES that have produced refinements to the vaguely defined DCMES categories. However, when these initiatives do not add additional elements, they are limited to a certain extent due to the strong requirement that the semantics of DCMES may not be extended.

The Dublin Core metadata initiative did not deal with structural embedding and implementation until very recently. The Architecture Working Group recently came up with suggestions of how to implement DCMES elements with the help of XML and RDF. Their suggestions indicate the confusion in the community that arose through the uncontrolled extension of the DCMES. For matters of easiness it is suggested that “refinements of elements are elements in their own” which would indicate a change in attitude in the DC community. So, implementation considerations could lead to another phase of changes.

It is widely accepted that the DCMES designers made a wise decision in defining a simple, flat set where the semantics of the elements are vaguely specified. This will allow the data manager to describe a large variety of resources on a shallow level and the general web-user to improve the discovery of these resources compared to what would be possible at present with the usual search engines. There is no doubt that DCMES is currently the most important standard for the simple description of electronically available, simply structured resources. However, the metaphor of “pidginization” on the one hand and “creolization” on the other hand as used by Baker indicates the inherent problems of a metadata set such as DCMES.



### Metadata Sets for Language Resources

On a low level of detail the **Dublin Core** metadata element set could be seen as a set with help of which language resources can be described. However, DCMES has severe limitations if more granularity or detail is required. For example there is no distinction between the language a document is written in and the language a document is about. In a typical linguistic document containing for example annotations, one will have much information written in English although the language under investigation might be a Maya language such as Tzeltal. DCMES has an element “DC:Language” which is meant to codify only the language something is written in. Another element “DC:Subject” is used to describe the document is about. However, DC does not make further clarifying statements what this can mean. So one could use this element with an appropriate sub-element or refinement to describe the language the document is about.

This deficit with respect to language resources was correctly identified by the *OLAC* (Open Language Archives Community) motivators from LDC and SIL. In the DC community currently two mechanisms are used for extensions: refinements and sub-elements, both having highly overlapping semantics. Based on an inquiry about user needs, OLAC came to the conclusion to add a number of sub-elements. One is for example the language a document is about and another is the linguistic data type the metadata description refers to. Further, to narrow down the vague semantics of the DCMES elements OLAC defined a number of refinements of the element semantics. This makes the OLAC set which was derived from DCMES much more usable for describing language resources. OLAC makes statements that its metadata set can be used for all type of language resources, even to describe tools for language resources and best practice advice. OLAC started to define a number of controlled vocabularies that should be used within their set. Due to its approach, OLAC wants to provide a metadata set that is useful for the whole language resource community. They also see their metadata set as an umbrella to achieve interoperability on metadata level for the language resource community.

While OLAC started from the existing DCMES set to define a metadata set for language resources, the *ISLE Metadata Initiative* (IMDI) voted for another way. It started with an inventory of existing metadata practice in the language resource domain and a discussion about the actual needs of the users, especially when dealing with multimedia/multimodal resources. IMDI decided to only describe the major types of language resources (corpora, lexica) by metadata, since their structure and content has been analyzed in sufficient detail. Other data types such as grammar descriptions and field notes are less well described. Therefore, it is not yet understood what the requirements are to describe them to facilitate discovery. The IMDI set contains much more detail compared to DCMES, allows to enter descriptions at many levels and to specify the language these descriptions are in. In contrast to the flat DCMES specifications (just a list of elements) the IMDI set comes along with structure to be able to express that for example different participants have different attributes such as age and sex. It also allows the user to maintain the strong relation between related resources such as media files, annotation files and where available the sources (for example the original tape) implicitly. The IMDI initiative had the task to deliver a complete environment; therefore also several controlled vocabularies are supported. Knowing that we miss experience much degree of flexibility is built in into the IMDI set. Since the IMDI initiative accepts the DC/OLAC role for interoperability with the more general OAI community, a mapping was implemented which is not without information loss.

### **Strongly Related Metadata Initiatives**

The media and film community seems to broadly support the development of the *MPEG7* standard which is also adopted by ISO and IEC. MPEG7 was started by looking at existing standards such as SMPTE and the emerging requirements of the community. The result was an exhaustive element set combining both, suggestions for annotating film productions but also for creating metadata descriptions. The focus in film industry is clearly to support the production process, i.e. also annotate movies with low-level features such as for example “scene change”. In the object oriented MPEG4 decoding scenario MPEG7 is intended to support the query, selection and filtering process of the user such that he can easily assemble clips and other information to new personalized presentations. MPEG7 has defined its own Description Definition Language to define Descriptors and Description Schemes that is based on XML. Many descriptors and description schemes have already been defined including a “linguistic” one defining how linguistic phenomena can be encoded. MPEG7 contains many descriptors that match with the definition of metadata in this note. In the Harmony project MPEG7 defined a very restrictive mapping of its metadata elements to Dublin Core specifically to not extend the semantics of DCMES. Although MPEG7 is not particularly designed for the view on multimedia language resources people from the linguistic community have, MPEG7 will have strong impacts on this community, since it gives the possibility to define suitable schemes for sub-communities.

Under the umbrella of IEEE the *LOM* (Learning Object Metadata) standard is in the process of being defined. It makes use on the knowledge gathered in DCMI, but proposes an exhaustive set of elements that is necessary for the sufficient description of learning objects. LOM not only specifies the set of elements together with constraints for order, value range and basic data type, but also groups the elements into categories. Similar to IMDI implicit structure is defined by including aggregate and simple elements. Also the LOM initiative started to define controlled vocabularies for a number of data elements.

In the area of Content Management the *CEI/IEC 82045-1* standard has been established. It is a joint effort (JWG15) for a metadata element set from ISO TC10 and IEC SC3B. The proposal speaks about management data as data about the content of an electronic or paper document, necessary to manage it in an Electronic Document Management System. Based on a broad analysis of possible documents and document collections including their various types of relationships during their life-time, an exhaustive metadata set is developed. Also here the many elements are grouped together in a hierarchy for intelligibility reasons.

### **Repository Initiatives**

Beyond those initiatives that defined standards such as described above, there are also initiatives that are building up relevant repositories in the linguistic domain covering metadata. Here, we want to refer to initiatives such as the *DFKI Tool Repository*, the *COLLATE* Project at DFKI and the *TELRI* initiative.

The first two repositories will be merged, since *COLLATE* is intended to become a large collection of information about people, projects, initiatives, publications and tools in the area of especially language engineering. The tool repository is built upon a well-documented taxonomy that is ready to be formulated in terms of a metadata element set for tools. Other information in *COLLATE* is not yet so much subject of specific metadata description standards, although information about people can be formulated according to some norm which is also used by other initiatives such as *IMDI*. The *TELRI* initiative is intended to build up a resource database, but wants to combine this also with information about tools. Until now there is no intention to use formal methods, and also the advice about tools will be given by personal information.

### Terminology Initiatives

Initiatives for standardized and open terminology repositories (monolingual and multilingual) are of extreme relevance since they will contain the definitions of terminology units, i.e. concepts and the corresponding terms in the various languages. With respect to language resources a number of initiatives and standards have to be mentioned. *ISO 12620* has defined a number of types of data items - where each type is called a data category. The types are grouped into 4 major categories and serve to describe terms: (1) The first category covers the data category "term" itself; (2) The second covers all term-related information which have to be associated with a term such as usage and etymological information; (3) The third covers descriptive information which describes the meaning and relates the corresponding concept entry to other concepts; (4) The fourth covers the typical administrative information. *ISO 12620* does not specify the structure of term entries, i.e. nothing is said about the structural embedding of the different data items. Similar statements can be made for the *OLIF2* proposal that lists a number of data categories occurring in lexica.

*ISO (FDIS) 12200* is a concept-based interchange format for terminology databases in SGML described with the help of a DTD. *MARTIF* makes use of the data categories defined in 12620 without restricting data categories to the about 150 defined within 12620. *MARTIF* documents have global information (typical header information), a set of concept entries (body) and a set of references to shared documents. *MARTIF* separates into negotiated and blind *MARTIF*. While the first describes a flexible interchange format that two or more partners can agree upon, the latter refers to a pre-defined standard that everyone participating will accept blindly.

*CLS* is a framework to define the structure and content of terminology databases and the references that can occur. *CLS* is very much inline with *ISO 12620* and *12200*. *SALT* can be seen as a very recent initiative to extend the terminology work to lexica, in particular those used in the translation business. This step seems to be a natural one since term information is very much related to linguistic information in lexica. *SALT* therefore wants mainly to test and define an XML-based lexicon/terminology interchange format called *XLT* and to provide tools such as an editor that allows the user to create *RDF* (Resource Description Framework) descriptions for complex terminology units.

### Metadata Integration Initiatives

In this chapter we want to briefly discuss integration mechanisms for metadata proposals as far as they are known to us. We can distinguish a few different approaches where integration is an objective but on various unrelated levels:

- The *OAI* (Open Archives Initiative) approach is focusing on defining a searchable domain for metadata such that services can be built on top of search engines using structured metadata information. A simple protocol for metadata harvesting was defined. *OAI* is offering the means to harvest metadata records by service providers and requires that the data providers at least offer the records with *DC* elements. The implementation of any mapping between a domain specific element set to *DC* is left to the data provider.
- The *IMDI* approach is focusing on the management of language resources that includes both aspects: (1) creating a searchable domain as indicated above and (2) creating a linked domain of metadata descriptions which is browsable and which allows to include many different data types. *IMDI* therefore makes use of simple mechanisms to create an integrated metadata domain for browsing. For search it makes use of distributed databases accessed through the *http* protocol.
- The *COLLATE* initiative wants to establish linked *html* pages with various types of meta information in the area of language resources and use advanced *IE* technology to automatically create relational links between elements of the documents included.
- The *INTERA* project wants to integrate the two complementary types of repositories (resource and tool repository) to facilitate the selection and execution of tools on chosen resources by interacting agents.
- The *E-Meld* initiative wants to act as service provider for metadata in the area of language resources. In accordance to the *OLAC* standard it intends to combine metadata about resources, tools and advice and allows the user to combine the three types of resources by manual intervention. Also *OLAC* makes use of the *OAI* harvesting protocol.

## **RDF**

Due to its assumed relevance for the future *RDF* (Resource Description Framework) is mentioned separately. The RDF initiative by W3C is focusing on defining a common framework for complex metadata scenarios. It allows the designers to define data categories (simple or complex categories), their constraints and controlled vocabularies, and in particular the relations within a metadata set or between different metadata sets. Due to its goals RDF will open the gate to define metadata element sets such that they can play a role in the emerging Semantic Web. In the Semantic Web agents will operate on the element definitions and relations as defined in open, machine-readable repositories.

## **5. State and Challenges for the Future**

The creation of web-accessible metadata descriptions to facilitate the management and the discovery of language resources is a comparatively new concept. All initiatives discussed are relatively young and have undergone more or less a highly dynamic phase. Therefore, it is too early to draw conclusions. A number of relevant questions such as “What are the typical queries the different user groups will input?” or “What are the widely agreed elements and controlled vocabularies?” cannot be answered yet. Therefore, it is good that the mentioned multiple initiatives were started. Also it seems that we are approaching a situation with a critical mass of resources described by metadata. It may motivate other groups and individuals to join. The community can gather experiences with metadata and different approaches such that in a few years conclusions can be drawn.

The variety of approaches and initiatives can at first glance be seen as a disadvantage. However, it was already foreseen by the driving forces behind Dublin Core that a scenario with many different metadata approaches will emerge, since every community and even sub-community has different views about real entities and that the multiple views should not be integrated per se to one complex description standard. Only the experience will show which concepts are flexible and stable enough. Having foreseen this scenario with multiple approaches and initiatives they also started a discussion about how to achieve interoperability. As already mentioned Dublin Core seems to be the accepted pidgin set by most of the initiatives such that mappings of their element sets to DC categories are provided. But these mappings imply the loss of information or bear the danger of an increased creolization of the DC set.

The emergence of the Resource Description Framework and the elaborations about an ABC model indicate possible more advanced solutions for the interoperability problem. RDF imposes formal and machine-readable structure on top of XML to support consistent representation of semantic relations. Even more exhaustive standards such as DAML/OIL will add further possibilities to express semantics. 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. Such a network of interrelated machine-readable metadata components opens the view to the Semantic Web introduced recently by T. Berners-Lee. Here intelligent agents use the term definitions and relations stored in the web to execute smart searches and other tasks for the user.

Another view to metadata for future scenarios is that it will allow to blindly executing operations as they are used for example in Information Extraction. The well-known GATE system defines a framework where different IE components can be executed in a chain reusing management information. Each component can add the necessary information to the metadata description. Amongst other purposes this information indicates which NLP components can be executed next and where these components can find the relevant information in a distributed scenario.

Summarizing we can say that a number of metadata initiatives are maturing and that the infrastructures they propose are more and more accepted by the communities. Often they propose a mapping to the simple Dublin Core set as a first basis for interoperability. This will allow us to gather broad experience with various approaches. This experience will be necessary to discuss and design the framework we need to realize the metadata infrastructure for the Semantic Web.

## **6. Construction of WI-2**

The ISO working item on metadata descriptions about language resources has to be constructed to include all initiatives and projects which are clearly devoted to dealing with formal metadata and which have an interest in using metadata descriptions. The following initiatives are relevant in this respect and have to be in the list of official partners:

- TEI: as initiative having worked extensively in defining structures of textual resources
- DC: as most important metadata initiative world-wide with a claim for general coverage and interoperability

- OLAC: as the DC-based initiative in the domain of language resources
- IMDI: as the initiative in the domain of language resources covering more detailed descriptions
- MPEG7: as a highly relevant initiative in a closely related domain
- IMS/LOM: also as a highly relevant initiative in a closely related domain
- IEC 82045-1: also as a highly relevant initiative in a closely related domain
- COLLATE: as an initiative gathering much data in the domain of language resources almost ready to define formal metadata sets
- Terminology Initiatives: as initiatives which know much about the definition of data categories
- RDF/W3C: as an initiative which has the experts to show the way from metadata to the Semantic Web

The following contacts have already been established at an official level: OLAC, IMDI, MPEG7, LOM, IEC 82045-1 and COLLATE. Contacts to W3C, TEI and DC have to be renewed.

Further, we have to include expert users from the different linguistic sub disciplines working with metadata or having insights in metadata requirements:

- corpus and field linguistics
- language engineering
  - text-based work
  - multimodal work
- artificial intelligence
- phonetics
- psycholinguistics

Amongst these there should be experts for corpora, lexica and other data types being used in the domain of language resources.

## **7. Concrete Steps**

First, the official statements are repeated. The TC37/SC4 meeting initiated a task force and specified the following:

- as focus:
  - produce an overview about existing projects/initiatives and monitor its usage
  - link with activities of the emerging Semantic Web
- as tasks:
  - provide a clear picture of the needs of the other WGs
  - identify the experts
  - draft a requirement document until the end of 2002

The TC37/SC4 resolution document adds another point explicitly: create a basic paper on goals and views.

This paper is seen as a basic paper on goals and views. It also includes an overview about relevant initiatives, makes first statements about the directions the Semantic Web may take and describes from which initiatives and fields experts should be invited to build the task-force. The task of WI-2 can be split into two major phases: (1) In the current phase metadata initiatives worked out excellent proposals that are in operation right now. ISO should gather the experiences made with these approaches. (2) Based on the experiences and a requirement analysis WI-2 should work out proposals that meet future needs.

In phase 1 WI-2 should carry out the following concrete steps:

- 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.

It will require more preparation and discussions to work out the tasks for the future.