Workshop on Meta-Descriptions for Multi-media Language Resources

LREC Preconference Workshop

Summary

P. Wittenburg, H. Brugman, D. Broeder

Contributions

First the White Paper of the meta-description initiative within the EAGLES/ISLE project was presented by Peter Wittenburg. It argues that it is time to create a structured sub-space for the language resource community in the Internet to easily locate resources of interest. This will be achieved by describing the resources with meta-descriptions (header information), making them available for structured searches and by using them to create browsable hierarchies. The White Paper describes the problems to be solved such as defining a widely accepted set of meta-elements and describes an organisational structure to achieve that. Then Henry Thompson gave his view about meta-data. His great vision is that all data is open data in the Internet and that it is machine exploitable. In this scenario XML is one of the key components for representing "simple tree-structured" documents. Finally he argued that we need many projects, since we still don't have a stable "ontology" of the field. A key for the success of meta-descriptions will be the control of its quality. Steven Berman discussed the requirements with respect to meta-data from the point of searching. He made a distinction between the needs for local information and web-based information, since having searching agents crawling through the web to find hits is still a very expensive operation. He also suggests to come to a meta-data standard where meta-data is available as attribute-value pairs.

After these more general talks three talks were given which presented the ideas from a users perspective. Caroline Wilners described the corpus related work at Lund university and argued that even for internal purposes the availability of a browsable&searchable universe of meta-descriptions would help the researchers a lot. The current practice is that no one knows exactly which corpora are available and in what status they are. Nelleke Oostdijk explained why the Dutch National Corpus project relies on structured meta-descriptions to organize the project's data and allow better access to it for arbitrary users. She especially stressed the need of flexibility with respect to meta-descriptions. Pirkko Suihkonen refered to the work at Helsinki university where a web-site was setup to help interested people to get an overview about available corpora. She also presented a highly detailed list of meta-data categories and the meta-elements she needs to describe the resources at Helsinki university and MPI for evolutionary Anthropology.

Daan Broeder presented the meta-description project at the MPI for Psycholinguistics. A unified meta-scheme based on XML syntax is the basis for describing the many resources and for preventing a chaotic situation where only individuals know how to access the resources. He also explained what kind of tools were programmed to create meta-descriptions and browse through the universe of such descriptions. Finally, Khalid Choukri discussed the role of a resource agency like ELRA in distributing language resources and how this role may change over time. Internet will have its great impact, but still existing channels of accessing data via the ELRA catalogue and media distribution will the preferred method by many users.

Discussion

The discussion after the talks and at the end of the session resulted in a number of interesting points:

- Meta-descriptions will only be accepted when a high quality is guaranteed.
- Some people or institutions urgently need methods to prevent a complete chaos where only few individuals know about the state of corpus projects and the ways to access them. In these institutions typically many resources are created continuously.
- Many meta-description related aspects are highly dynamic, i.e. we will need several attempts and projects to fully understand the problems, unify the terminology, and come to a stable state.
- Some argue that it is better to not separate meta-descriptions and annotations. One reason is that the content of the annotations might change such that meta-data is effected. When meta-data and content data is separated it might be difficult to keep the meta-data up-to-date. Another reason may be the enhanced possibilities of search operations which could combine header and body search.
- The meta-descriptions must have a mechanism to allow flexible extensions for sub-communities. The problem with such extensions, however, is how to allow the search engine operate on them and how to inform the user of their existence and meaning.

- All resources should be openly accessible in the Internet.
- How to prevent an endless discussion about meta-elements?
- Are other initiatives such as RDF from W3C of any relevance for the meta-project?

Summary Statement

- We can identify an extreme increase in the number of language resources being produced world-wide. We urgently need ways to capture the knowledge about their content and construction and to make them browsable and searchable by the interested community. Some users from well-known research institutions have expressed their wish to start the meta-project and soon have a description standard available and tools operating on them. In companies working with many resources it is self-evident to have a database which describes them.
- The community is sceptical whether we will achieve the goal to have all resources freely available in the Internet rather soon. There are too many obstacles which will limit the general accessibility of the resources themselves. However, meta-descriptions could be openly available. In fact, the Talkbank project designed an elaborated access right system which may be taken as an indicator of how sensitive these aspects are.
- Although the authors see the problem which can occur when separating meta- and content information, there are 5 reasons which advise us to go ahead with the meta-description project:
    - There are many resources in native formats such as CHAT. It cannot be seen how all this data will soon be converted to XML-based formats. This means that there is no such hierarchically structured document describing meta-data and content. Separate meta-descriptions can easily be created based on the header information.
    - As already mentioned many resources will not be freely available on the net. Nevertheless, it is very useful for the community to know whether there are some available with certain characteristica and whom to contact to get access.
    - Searching in a web-based meta-universe will be much simpler and much less compute intensive than searching in a universe of the resources themselves.
    - There is an extremely high pressure to start creating a standard for meta-descriptions independent of the question whether they are integrated with the content or not.
    - There is no special problem to not hook up complete resources to the meta-universe, if the meta-description schema is identical and if the tools can cope with this. On the other hand it is a simple operation to extract meta-data from a complete document and integrate it in the meta-universe.
- If separation is done, of course, one has to set up a scheme which allows the provider to automatically adapt the meta-descriptions after the content was changed. Since the meta-descriptions are part of a distributed scheme there is no reason to not maintain them at the places where the resources themselves are stored.
- With respect to combined header and body searches methods can be imagined to allow these even in case of separate meta-descriptions.
- The quality assurance needs an appropriate organizational approach. No one may be allowed to hook up meta-descriptions to the universe without quality check. There has to be a clear system of authorization.
- It was clear that the structure of the meta-descriptions has to cope with flexibility and dynamics. The right technical mechanisms have to be worked out within the project.
- Continuously analysing the progress of other initiatives such as XML-schemas, RDF, DC, MPEG7 etc is a must. When it is possible to join or to take profit from these initiatives then the project should do.
- Agencies such as ELRA will be needed to control the quality of the meta-descriptions and to help integration and usage.

As a result of the workshop a Steering Board and a Technical Board for this meta-initiative could be setup and a the SB had its first meeting. An Advisory Board is in the process of being setup.

Comments and questions should be addressed to ISLE@mpi.nl