Meta-Description for Language Resources

EAGLES/ISLE

A Proposal for a Meta Description Standard for Language Resources

P. Wittenburg, D. Broeder, B. Sloman

18.5.2000

1. Goals

This white paper is intended to describe the work that has to be carried out in the EAGLES/ISLE sub-project "Standard for Meta-Descriptions for Language Resources", and to serve as a basis for discussion within the language resources community. The aim of this EAGLES/ISLE project is to improve the accessibility/availability of Language Resources (LR) on the Internet. We propose to achieve this by creating a browsable and searchable universe of meta-descriptions similar to those devised by other communities on the Internet. More and more language resources are being create across the world, and if these resources were tagged with the appropriate meta-data, researchers and companies would find it easier to discover the specific resources they need.

Meta-Data is data about data. A Meta-Description is the structured set of meta-data, which describes a certain language resource or a group of such resources in a way that it is meaningful to the user community. The universe of meta-descriptions should cover all the resources that the users might find useful, and the descriptions should be specific enough to allow the users to find precisely those resources that are relevant to their specific problem. The meta-universe of descriptions should be open to everyone, although access to the resources themselves can be restricted.

This White Paper is aimed at identifying and specifying the concepts to be described and will include the bare minimum on the format of the descriptions. Detailed formalisms can only be discussed once the basic concepts have been defined, but the project must eventually propose a specific standard.

2. Problem Description

Language Resources are collections of data representing examples of language being used, either directly, as in corpora, or as derived data, as in lexicons and ontologies. Fundamental and applied linguistic research has a long history of generating and using text-based language resources, and more recently multi-media language resources have been exploited in linguistics and related areas such as sign language, anthropology, computer linguistics, artificial intelligence, phonetics, psychology, speech recognition, multi-modal research and man-machine interface design. They are used for a variety of purposes. Linguists use them to create and test new linguistic hypotheses; speech recognition engineers use them to test speech recognition devices and to set recognition parameters. Increasing amounts of money are being spent on creating new language resources and extending language resources to combine a variety of inputs (sound, video, eye tracking,…) and to incorporate multi-modal annotations.

People have always referred to such language resources in terms of basic global characteristics such as "the resource includes speech by a 6 year old male Tamil speaker born in a farming environment" or "this resource includes pointing gestures and speech utterances recorded when people were asked for directions to the railway station". We call this kind of data, which briefly summarises or characterises the content of the resource, its meta-description. Most language resources include this data in "headers" [1]. These are either part of the resource itself or exist as separate files in a corpus-specific format. Every project defined its own header structure and content, appropriate to the goals of the project. Special tools or simple ASCII editors were used to display the data.

The development of the World Wide Web, with its linked web pages, provides new opportunities. We envisage a universe of linked meta-descriptions offering the interested community information about existing language resources. This universe should be accessible via the Internet with appropriate tools for browsing and searching. Such a system could save researchers and industry a lot of time in locating the appropriate LRs.

We are sure that the general addition of Meta-Descriptions to Web-accessible data will eventually revolutionise the way the Internet is used. Thus the language resource community should test these new mechanisms at an early stage,

both to confirm that the mechanisms and procedures now becoming available are capable of providing the services required by the language resources community, and to select the specific mechanisms and procedures which best match the needs of the community.

Making a universe of linked meta-descriptions available for browse and search operations (resource discovery) requires the development of a standard for the structure and semantics of these meta-descriptions. But language resources vary considerably and there is a heterogeneous user community, so we have to ask ourselves whether the requirements for handling such a variety of applications can be captured within a single standard. We have to look at the ways other communities have handled similar problems. The Dublin-Core standard [2] defined by the librarian's community would seem to be a relevant example. If we can exploit the experience built up in the development of this and other evolving standards, it should be possible to propose an ISLE-Meta-Description standard within two years.

3. Previous Work in the Language Resource Community

The previous work in the language resource community related to Meta-Descriptions can be grouped into four phases:

(1) header information in proprietary formats,

(2) header information as TEI-tagged meta-data embedded in the basic data file.

(3) header information as untagged, human-readable information in web-accessible hyperlink structures,

(4) tagged Meta-Descriptions in web-accessible hyperlink structures.

Reviews of language resources like the Bird/Liberman annotation web page and the MATE project resource reviewdo not pay much attention to header data meta-descriptions. Frequently it isn't even mentioned. The famous CHILDES [3] project developed the CHAT format in the mid-1980s and is an early example of the use of headers in a proprietary file formats to code meta information, such as the names, age, and sex of the persons involved, the language spoken and so forth. Special tools gave separate access to such header information, which was incorporated within the annotation files. The whole corpus was organised into a well-defined directory tree such that a person who knew the corpus could easily find their way to the data of interest. A great many people have adopted this format and large corpora of child speech exist. Other projects have encoded and stored meta information in their own - usually similar - formats.

The emergence of large text corpora created a requirement for a standard encoding format. The TEI text encoding initiative, which ran from 1990 to 1994 (in the first instance), defined header information standards for various types of text resources and adopted the SGML mark-up language as a unifying format. However, header information still was embedded within the annotation files.

This reflects an approach in which the document structure is seen as a hierarchy, with the head of the hierarchy holding encoded information which refers to the document as a whole. The TEI initiative was influential, as can be seen from the CES [4] and MATE [5] introductory web-pages, and its standard for header information was widely accepted in the language resource community. Nevertheless, the limitations of the TEI standard soon became apparent, and some of the subsequent projects used their own extensions of the basic TEI approach. An example is the BBAW digital dictionary of German. Although TEI-compliant, it adds specialised header information specific to the needs of the project. To make the header information rapidly available to a special search engine, the header information was segregated into separate files. The new TalkBank [6] project, funded by the American National Science Foundation, introduces their CODON encoding architecture which will XMLbased. It can be assumed that TEI-based header information may be represented in CODON.

The emergence of the World-Wide-Web encouraged many projects to make their corpora accessible over the

Internet. Sites were set up which allowed the user to browse through a network (mostly hierarchies organised by language groups) of linked web-pages. The LDC [7] and ELRA [8] agencies are two substantial examples. Their catalogues are available on-line (though LDC does require that potential users first register, and pay a substantial registration fee) and offer lists of resources sorted along a few different dimensions, such as resource type, and year of creation. Each entry in these lists offers very limited information - essentially a page of meta-data elements with some explanatory text, mostly associated with the language used, since this is usually the primary category. A search engine is available which operates on these meta descriptions.

Other more project-related examples of this approach are the Australian AIATSIS [9] electronic archive, and the University of Helsinki Language [10] Corpus Server which offers log-in access to computer corpora of more than 50 languages. The nodes in these networks of web-pages often include some meta data, albeit in non-standardised and frequently inconsistent formats. It is effectively impossible to search the universe of web-pages for these specific pages on the basis of the meta data they contain.

In 1998 by the Max Plank Institute for Psycholinguistics proposed Browsable Corpus scheme [11]to organize the rapidly enlarging (and essentially chaotic) body of resources that were being generated by the many field workers and linguists working on language acquisition. It allowed the researcher on an intranet to browse and search a local universe of meta-descriptions to find relevant data, then it automatically selected and started the tool appropriate to the specific corpus selected.

The British ICE project [12] developed a similar scheme based on ICECUP tool. The latest version of this tool, which was issued recently, offers access to a corpus map, where every node in the hierarchy is associated with textual descriptions which can be read directly. This relies on a unified ICE format for the descriptions.

These schemes form the basis of this White Paper.

4. Requirements of the LR Community

We have to produce definitions of the terms "language resource" and "language resource community".

"Language resources" are those databases which primarily document communicative acts of humans by some form of recording and/or descriptions, both directly as in corpora, or at higher levels of abstraction in lexicons and ontologies. Similar resources have been created, amongst others, by researchers working with chimpanzees where the communicative acts are studied and annotated. This work doesn't involve language as it is usually defined so it lies outside the scope of this paper, but there should be enough flexibility (vocabulary, structure) such that this sort of work could be included. Human communication is either verbal or gestural; i.e. the basic material is either an audio recording or a video recording including audio tracks. When recording dialogue or discussion we can be confronted with several audio and video tracks. We have to make allowance for situations where other data is recorded by non-video techniques such as eye- or body-movement, or where brain images, EEG signals or articulation data are also recorded.

These basic recordings (time series) are supplemented by various annotations. These are tiers of manually or automatically generated textual descriptions. The manually generated tiers can be either free text or code generated from constrained input sets. These annotations can made for a wide variety of purposes, largely dependent on the needs of the discipline. Linguists usually add layers of orthographic transcription, English translation, morphological coding and syntactical coding, and may abstract higher level descriptions like lexicons and ontologies. In disciplines where body movements in dialogues are studied, many tiers of annotation describing the gestures of the subjects will be added. Traditional language resources were restricted to textual descriptions, originally because they preceded reliable portable sound recorders, and more recently because it was too expensive to present digitized versions of the original recording material. Some modern language resources still include only textual material because they document written language usage, or concentrate on higher level abstractions such as words or sentences, as in lexicons and ontologies, amongst other secondary resources.

The "language resource community" to be addressed by this white paper, is the group of researchers and developers working with such language resources. These can be either researchers using such resources for theorizing or testing new hypotheses, or technology developers who use such resources to train their statistical recognition machinery. It is not assumed that the language resource community is a monolithic whole, where everybody has the same interests. We easily can identify such sub-communities as anthropologists who would like to be able to structure their data-bases around geographic references which are of little or no interest to other sub-communities. There will be other sub-communities with equally specific interests.

Granting that this is an adequate definition of language resources and the language resources community, we want to make a first attempt at the requirements for meta-descriptions. We assume that the community is interested in being easily able to find out whether resources with certain characteristics are available, and what is required to access these resources.

When looking for usable resources the users are not interested in a time consuming analysis of their contents, but they are interested in meta-data which describes the form and content. A researcher working on longitudinal studies for Nordic languages might be interested in finding those resources which contain recordings of Scandinavian children taken every two years during the first 15 years of life. A producer of machinery which has to detect pointing gestures automatically, might be interested in finding those resources which contain video recordings where gestures are recorded and pointing is coded. In most cases the text content of the language resource doesn't lend itself to being searched for this sort of meta-data - unstructured header data is too difficult to distinguish from annotated content.

What we therefore need is a universe of meta-descriptions of such resources which contain the relevant descriptors and pointers to the resources and to the institution that owns them. What we also need is a web portal, which can open this universe to the interested user. Finding the descriptors via the web can either be done by browsing the linked meta-descriptions, or by formulating and executing queries directed at a search engine, or by applying a mix of those two strategies.

Efficient browsing is dependent on the availability of intuitively understandable hierarchies formed by grouping resources together that share certain (meta-data) characteristics. Experience shows that such hierarchies are difficult if not impossible to establish in large and anonymous groups. Nevertheless, it is useful to be able to navigate visual representations of such search spaces. If the domain is not too broad, search strategies which can operate on exact assertions can be more powerful. Thus efficient searching requires a data-space which uses very well defined categories, with precisely defined attributes which are restricted to a bounded range of values, which means that the meta-descriptions should be based on well-defined assertions, expressed in a common syntax.

Another requirement for efficient searching is, of course, wide agreement about the categories that describe the relevant aspects of the resources for the whole community. The vocabulary has to be specified and the semantics have to be laid down in accessible documents. Finding such categories and specifying their semantics is a time-consuming operation in heterogeneous communities. From comparable initiatives it is well known that the vocabulary has to be small if endless discussion is to be avoided. Restricting the vocabulary implies that certain fine-grained distinctions will not be available in the meta-description universe, which in turn requires that enough flexibility has to be available for sub-communities to add specific descriptions they might find important.

This white paper will not go into further detail about the vocabulary to be used. To chose the vocabulary a network of interested people has to be formed and a structure of discussion has to be described. In chapter 2 we have already referred to the descriptors used in the traditional headers and to the meta-description format for the Max-Planck-Institute for Psycholinguistics. In chapter 6 these aspects will be further discussed. What can be said at this moment is that for the EAGLES/ISLE initiative the only feasible approach is one where we restrict ourselves to describing "language resources" as required by the language resources community.

It would be advantageous to maintain compatibility with other meta-description mechanisms used on the World Wide Web. XML [13] is the accepted standard for the syntax of meta-descriptions and RDF [14] is seen as possibly

providing a framework to which the LR community could conform. This is described in more detail in the following chapter. We have to make sure that meta-data which is already available in the web - such as addresses of people or companies - can be re-used; i.e. intuitively understandable concepts should be shared with other communities, if the chosen meta-data mechanisms are compatible.

5. Work in W3C and related communities

The enormous increase in the number of web pages and the seemingly endless variety of information available on the Internet has made it necessary to think about new access strategies. Meta-descriptions are seen as a means to provide browsable structure on the Internet and as a means to define searchable spaces. According to T. Berners-Lee, one of the driving forces behind the World Wide Web, meta-data is machine understandable information about web-resources [15]. The architecture of meta-data is represented as a set of independent assertions. The PICS initiative was the first to define meta-data to allow parental supervision of the web-data children can access [16]. Companies such as Microsoft and Netscape identified similar needs and came up with the WebCollections [17] and MCF [18] proposals respectively. Librarians as a community followed by defining the Dublin-Core standard, which can be used to describe contents of Digital Libraries in the most general sense[2].

The Dublin-Core (DC) defines 15 core elements and describes their meaning in web-accessible documents. This limited set of core elements can be extended by attributes and qualifiers to give more space to describe specialised documents. The core elements contain specifiers for topics such as the creator or title of a document, or the language it is written in. XML was chosen as the formalism for the syntax of DC and for structuring the meta-documents. The core set was limited to 15 elements to achieve semantic interoperability and give general users a uniform and easy-to-understand description. The definition of the vocabulary and the semantics is still going on and seems to be consuming a great deal of work.

At the moment it seems that the DC initiative is divided: One faction wants to restrict specification to a narrow set of easily implemented descriptors and options, while the other faction wants to extend the meta-model so that it can describe all kinds of documents, including language resources.

Businessmen have organized a standard for web-accessible business cards called CARD [17].

This plethora of initiatives lead the W3C to define a general framework for meta-descriptions called the Resource Description Framework (RDF). The RDF initiative is a very interesting one since it offers a unifying framework such that several different communities can share meta-definitions. For example names of persons will be used in many communities for different purposes. In the LR community a name could define a person who created the syntax annotation of a certain corpus file. Searchers might be interested in the affiliation of this person. The LR community could specify their own coding for the affiliation of such a person or they might chose to rely on the work of the CARD community, since the CARD community's meta-descriptions associate persons with a number of characteristics which include affiliation. The RDF standard makes it possible to combine the meta-elements of various communities.

RDF uses the standard XML syntax. The vocabularies and the semantics are described by the various communities and RDF just offers a structural framework to bring these together. Name Spaces can be defined to refer to the definitions of various communities.

The MPEG7 standard initiative [19] uses the term "meta-description" as already mentioned. MPEG7 is to define a standard description of the structure and content of movies such that people can search them to find specific scenes and such that the MPEG decoder can process the meta-data stream in real time to do selections etc. MPEG7 therefore views all descriptions which are associated with a movie as meta-data. This includes elements which describe the whole movie as well as elements which describe a certain feature of a sequence of video frames. Since MPEG7 decided to use XML as the underlying syntax for their descriptions, it is not surprising that they have chosen this approach. XML allows the user to create a hierarchy of such descriptions, starting at the top with the whole film as a subject of description, and ending with descriptions of the individual frames where necessary.

Of course, it would be possible to automatically copy the elements describing the whole content into separate files and integrate them into a browsable universe. Thus, there isn't in principal any problem involved in integrating MPEG7-coded resources into a meta-universe. In the media world a movie is shot from a single "shooting" script, so the creation of the "meta-descriptions" is almost automatic, while in the language resource community the creation of meta-descriptions requires additional mental effort and can be seen as a separate intellectual step. This might explain why the language resources community differentiates header information - meta-data - from annotation, where the moving picture community lumps the two together.

6. Problems to be solved

This chapter is devoted to listing all the major problems, which have to be tackled and solved within the EAGLES/ISLE meta initiative. Some of these problems are inter-dependent, and the ordering of the list reflects this, but other orderings are possible.

6.1 Goal

The goals of the EAGLES/ISLE meta-initiative given above were originally set out in the proposals submitted to the EC, but they have to be accepted as appropriate by the language resources community before any significant work can be done.

There are two deliverables:

1. A proposed standard for meta-descriptions for Language Resources.

2. A showcase demonstrating what the proposed standard would look like and how it might be used.

There is no guarantee that the proposed standard would actually be used, but there is an obligation to publicise the proposal within the language resources community. The MPI expects to have to provide the showcase.

6.2 Influences

The EAGLES/ISLE work on a meta-standard has to reflect two influences; on the one hand the expectations and demands of the language resource community, which has to specify its requirements, and on the other the current spate of meta-data initiatives, especially those arising within the World Wide Web Consortium (W3C). The proposed standard will have to map the requirements of the languages resources community, ideally onto the structures being built under the aegis of the World Wide Web.

6.2 Scope of LR and Community

As mentioned above, we have to define the range of the Language Resources to be covered by our meta-initiative. This effectively determines the data which has to be covered by the meta-descriptions and the scope of the standard, and defines the community which we want to address. As already mentioned, we plan to adopt a restrictive - as opposed to an exhaustive - approach. This is dictated by the project funding, which is limited to two years, and the urgent need - within a part of the community - for a standard and the tools to implement it.

We have to identify the sub-communities which may require the inclusion of specific - possibly unique - data in the meta-elements to be used to characterize their Language Resources. There are several dimensions within which sub-communities can be identified. When we look at the types of language resource we can distinguish textual corpora, annotated corpora, multi-media corpora, lexicons, typology databases, grammar notes, notes about sound-systems, ontologies, and others. This may be the best place to start. Another dimension can be defined by the specific research or development needs of particular sub-communities. Anthropologists will want to include different descriptors for multi-media/multi-modal corpora than will people designing man machine interfaces.

6.3 Structure of Meta-Descriptions

XML seems to be a natural choice for the underlying syntax of the meta-descriptions. It has the necessary power to express the required structure and also will allow us to use an expanding range of available software for parsing and generating purposes. Probably there will be a core set of mandatory meta-data elements which will be appropriate for all types of language resources. The presence of the core set would identify a file of meta-descriptions as describing a language resource, and allow a basic browse and search tool to navigate this universe of meta-descriptions. There would also be a significant range of common elements which would be optional, because creating meta-descriptions is a time-consuming task and we would not want to burden the users with inserting unnecessary data. In order to achieve the desired flexibility to cope with idiosyncratic sub-communities and specialised classes of LR, there would be open extensions which can accommodate specialised vocabularies of meta-data elements.

The detailed discussion of the management of such a flexible system requires further study.

It is not as yet certain that the model proposed by RDF is suitable for our needs. MPEG7 has apparently rejected it as incapable of handling multiple layers of annotation, but since we are not proposing to use RDF to encode annotation data, this doesn't seem to be a problem for us. Adopting existing standards has obvious advantages if they can provide the services required without importing too many unnecessary features, but there is no virtue in adopting an inappropriate standard.

6.4 Scope of Meta-Data

We have to define the scope of the meta-data; i.e. what type of solutions do we want to offer. It is clear that we want elements, which describe form and content of the related LR. Do we want to include elements

- which indicate the set of annotation tiers included
- which indicate standard or standards used to create the annotations (EAGLES …)
- which point to the original video tapes (which will probably never be used)
- which have some information relevant for certain tools
- which contain statements about rights, forms of accessibility, and payment
- others?

6.5 Meta-Data-Element Vocabulary

The definition of elements will take most of the time. The work will include bringing together specialists for the major dimensions to be identified (type of resources, sub-communities). Existing header file definitions and current practice at such institutions as the MPI have to be analyzed to identify descriptors which have already proven their usefulness. From these discussions the set of elements to be part of the standard proposal has to be defined. For this elements the semantics then have to be defined and published as a web-accessible document. Examples have to be created to make the semantics as explicit as possible to the normal users.

6.6 Meta-Data element mapping

Some meta-data elements are not to be directly accessible via the Internet. In many cases the names of subjects may

not be made public and here a kind of mapping mechanism must be used to replace the true name of the subjects by an alias that can only be resolved by a mapping file which is not publicly available.

6.7 Re-usage of meta-data element definitions from other communities

Assuming that a unifying mechanism such as RDF can be used, we have to check whether it makes sense to make use of elements already defined and used by other communities such as DC, CARD etc. These definitions have to be stable and must appear intuitive to the language community.

This sort of judgement seems to require access to the people who defined the elements for their various communities. The history of the Dublin Corpus initiative suggests that it can be difficult to just create definitions which mean the same thing to different members of the same community, let alone the members of different communities.

6.8 Requirements for tools

Ultimately we have to define requirements for tools to work on meta-descriptions.

There have to be meta-description **editors**, which help the naive user to enter the descriptions. These editors have to support the user by making the semantic descriptions of the elements available on request, and have to be flexible enough to cope with the structural flexibility required of the meta-description format.

We need suitable **browsers**, which understand the structure of the linked meta-description files and provide graphic support for the user during navigation. Since the standard browsers such as Netscape and Internet Explorer do not give us the desired functionality, we can either enhance their functionality by developing special applets or develop new proprietary browsers specific for the LR meta-descriptions. This might be necessary if - for example - we wished the application tool to be able to start to operate on a selected corpus file at the click of a mouse button.

We also need **search** tools, which can cope with the meta-description file structure and any meta-data elements taken over from other communities. The search tools have to use the links between the meta-descriptions and knowledge about available meta-descriptions efficiently. In cases where there are many meta-descriptions to be processed, optimization techniques such as caching pre-parsed meta-description files in a data-base can be required to keep response times within reasonable bounds.

6.8 Practicable Scenario

In our final documents we should describe a practicable scenario. This includes such topics as determining

- where to store the meta-descriptions
- ways to register and link the meta-descriptions
- ways to build browsable hierarchies
- ways to supervise the linking of new descriptions to the existing universe
- the requirements for centers which could establish and maintain such a universe

The Internet is growing dramatically and we need to understand how we can use it. As yet, nobody has tackled the topics mentioned above. We are faced with new challenges and must devise new solutions. The proposals which should emerge from our EAGLES/ISLE project are intended to be signposts on the route to establishing the sort of web we need.

7. Organization of the work

We need to take some organizational steps to get the initiative under way. In this chapter we suggest the levels and type of discussions we believe are necessary to achieve success.

Broad network of experts

First we have to establish a broad network of experts. These are people from the community who are committed in some way to contribute to the eventual goal. This network is open to every member of the community and all suggestions, questions etc will have to be taken up to be discussed. Every member of the network will be asked to persuade other experts from the discipline to contribute. We expect that the network will take the form of one or more closed user-groups (comparable with the Linguistics List, but without enough support to be accepted as regular news-group, distributed by news-servers), posting questions, proposals and responses. Contributions will be screened by a moderator to discourage off-topic input. The MPI will provide the server to display the contributions on the web and - in the first instance - a moderator to screen them. At regular intervals the members of the network will be informed about any qualitatively new information that has become available on the pages and will be asked for comments. Anyone will be able to subscribe to the user-group to receive e-mail copies of all contributions posted to one or more of the lists available on the web. The web-pages will be open to everyone. Public versions of documents to be discussed will be posted to the web-pages as they become available.

The web-pages will be divided amongst the dimensions mentioned (type of resources, sub-communities), if and when this becomes necessary.

Workshops & Conferences

The concept of a universe of meta-descriptions and the progress of work will be presented at workshops and conferences. All members of the network will be asked to help promoting this idea and to mention interesting platforms for discussions.

Steering Board

A small steering board will be set up consisting of committed people who are able - to some extent - to represent specific sub-communities and the different European areas. This board will have discuss the various proposals in detail, comment, and make decisions. Members of this board will have to actively participate in the public discussion process. This board has to meet at regular intervals and may ask other specialists for advice where necessary. The work of the Steering Board will be supported by a closed unmoderated email-list-server displayed on pass-word protected web-pages.

Technical Board

There may be a Technical Board to help the Steering Board in its decision making process. This board should comprise specialists who understand in detail the technologies involved and have some appreciation of the work going on in other similar communities. The TB will take over certain tasks defined by the SB and make detailed technical reports. The TB is not a fixed group, i.e. new members could be invited to join to contribute on specific topics. Despite this the size of the TB would have to be limited in the same way as the Steering Board (and for the same - financial- reason) . The chairman of the TB will be member of the SB as well. The work of the TB will be supported by an unmoderated closed email-list-server. The members of the TB will have access to the same protected web-site as the SB.

Official Documents

The publication of Open Documents which can be cited as official notes of the EAGLES/ISLE meta initiative will be the responsibility of the SB chairman. These documents will be discussed in the SB before publication, and there

will have to be a general agreement on basic issues, but the chairman has ultimate responsibility under the terms of the EAGLES/ISLE contract.

8. Procedure

It is proposed that this White Paper is initially discussed within the existing network of experts, to generate a document that can be presented and distributed at the LREC Workshop at the end of May 2000. We hope that it will be widely circulated in the potentially interested community, and will post it on the relevant mailing lists and user-groups. At the LREC workshop we will have a first open discussion about the concept of meta-descriptions for language resources, in the hope of getting additional support, and enlarging network of experts involved. The first Steering Board is to be presented at the LREC and perhaps a few members of the TB.

It could be that sub-communities will organize themselves and create documents. These would have to be discussed in the SB (and TB if necessary) and should be integrated into official documents where possible. If the SB does not agree with the positions of such a sub-community, the subject of dispute has to be submitted for discussion.

It is expected that closed workshops will be held at regular intervals, where the SB and members of the TB should be present and additional specialists may also be invited.

9. References

[1] http://www.mpi.nl/world/tg/lapp/esf/esf.html

[2] http://purl.org/DC/

[3] http://childes.psy.cmu.edu/

[4] CES http://www.cs.vassar.edu/CES/CES1.html

[5] MATE http://mate.mip.ou.dk

[6] TalkBank http://www.talkbank.org

[7] LDC http://www.ldc.upenn.edu

[8] ELRA http://www.icp.grenet.fr/ELRA/home.html

[9] AIATSIS http://www.aiatsis.gov.au

[10] http://www.ling.helsinki.fi/uhlcs/index.html

[11] http://www.mpi.nl/world/tg/lapp/browscorp/browscorp.html

[12] ICE-GB http://www.ucl.ac.uk/english-usage/ice-gb

[13] XML http://www.w3.org/XML

[14] RDF http://www.w3.org/RDF/

[15] Berners-Lee http://www.w3.org/People/Berners-Lee/

[16] PICS http://www.w3.org/PICS/

[17] MS WebCollections http://www.w3.org/TR/NOTE-XMLsubmit.html

[18] Netscape MCF http://www.w3.org/TR/NOTE-MCF-XML

[19] http://drogo.cselt.it/mpeg/standards.htm