WP2 View on Resources in ECHO

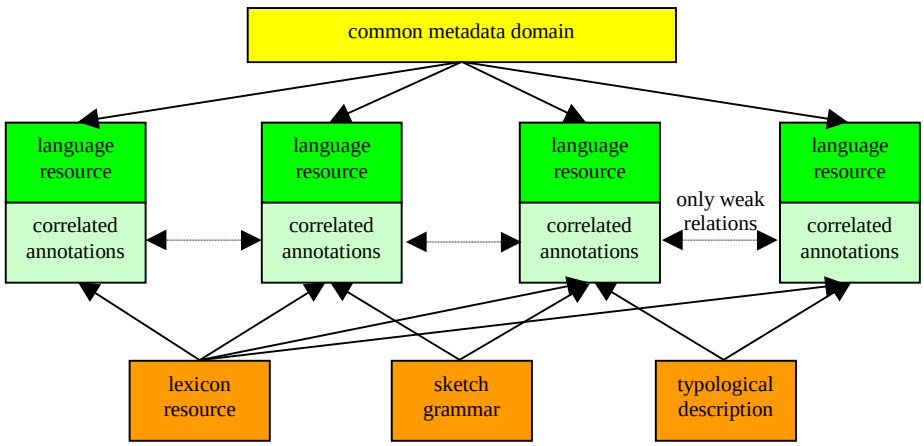
Peter Wittenburg
28.12.2002

On the basis of the resource overview one can identify a number of different views on resources between disciplines that all come together under the ECHO umbrella. They are described in the following, since they are the basis of understanding what a Common Technological Infrastructure could mean.

1. Language as Object or Medium

1.1 Language as Object

For the linguistic domain the recording is the basic unit of analysis and the relation between recording units is given by linguistic metadata such as common language, same speaker, age of speaker etc. The recording can be available as texts (where only a transcription is available), as sound or as video material. In the latter two cases mostly tightly coupled annotations are available at various linguistic levels. In general, there is no need for the linguists to establish links between the different resources except if one includes lexica, grammar descriptions etc. This kind of linguistic secondary data has references to various recordings and vice versa.



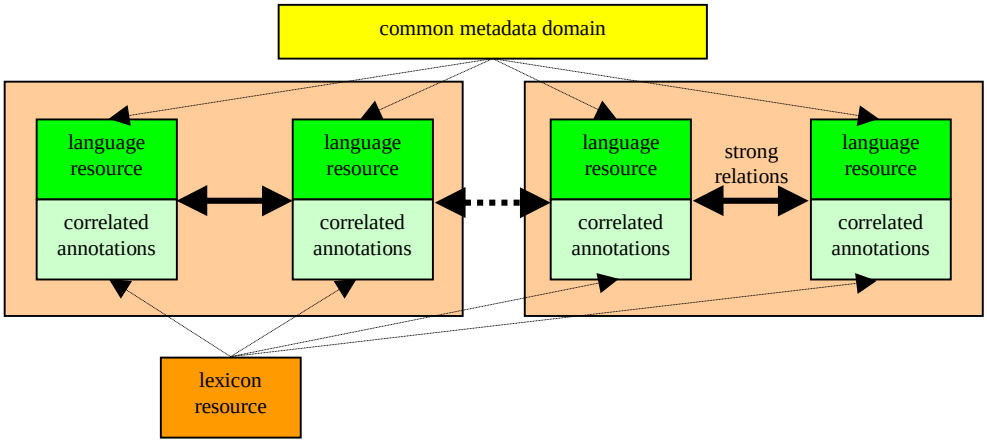
The domain knowledge is contained in the annotations and the derived data such as lexica, grammars, etc. Analyzing linguistic data often means finding related resources based on metadata, “looking” at the primary and secondary resources, compare instances etc. Resource clustering basically is often defined by projects, i.e. aspects that are not so relevant for later linguistic analysis.

In history of Arts a comparable paradigm can be found, since each individual arts object or groups of related object is subject of scientific interest and the relations between them are often documented in the metadata.

1.2 Language as Medium

For other domains language is a medium to convey descriptions of objects or ideas that are related in various ways. They want to semi-automatically establish the relations that are defined by the content of

the documents and not by metadata. Metadata identifying the singular resource plays a minor role, since the notion of the "singular resource" is not so clear and less important. Metadata could serve to identify the clusters of resources that exist. Original files are mostly images or texts.



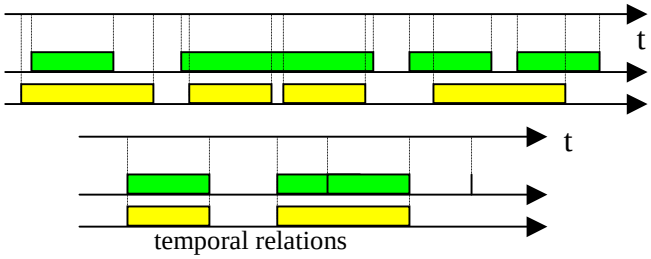
Such resources are clustered around certain thematic areas such as specific scientists. The relations between the resources of one cluster are very strong and important, those between clusters may be important as well dependent on the nature of the themes. The relations can be of many different sorts. Secondary resources such as lexica are basically only used to establish the relations via (semi-) automatic processes. Analyzing such data often means looking at the found relations and the corresponding resources.

2. Nature of Transcriptions and Annotations[1]

2.1 Time and Sequential Transcriptions and Annotations

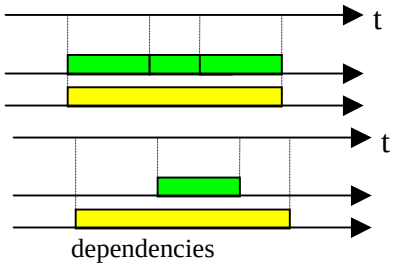
Assumed is a "basic raw resource" that has a linear one-dimensional extension such as texts from a book, texts typed over from a scanned manuscript page and transcriptions of a sound or movie file (includes the availability of several tiers dependent on the number of persons involved). Such texts are either linked with a time axis or just have an inherent sequential representation. Markup denotes a structural framework superposed on such texts mainly for referencing and identification purposes, but does not change the sequential dimension of the underlying text. In the case of texts copies of the "basic raw resource" could exist to superpose different types of structures on such sequences or to add extensions to it.

For the following consideration we assume either an existing reference with time dimension or a sequence of characters. Mostly both co-occur, since if for example a sound file exists the transciption(s) will only be aligned with the sound samples at the level of larger linguistic units such as utterances.

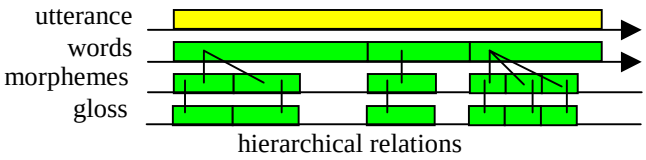


The two drawings show different types of temporal relations between some media file and two tiers of transcriptions. Chunks of characters are linked to the time line. The temporal relations between the two tiers in the first example are absolutely random as they occur whenever one can speak about independent channels (several speakers, gesture and speech, ...) being transcribed.

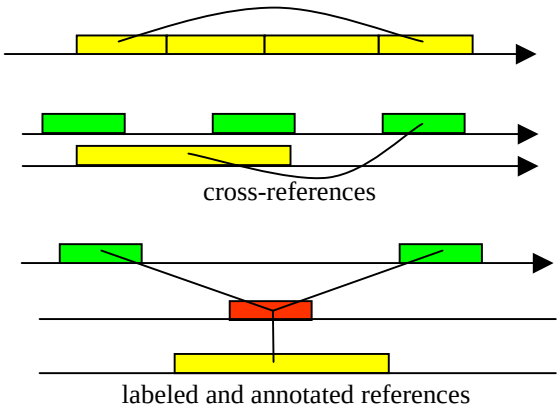
[1] The term "annotation" is used here in its narrower sense, i.e. texts associated with existing "basic texts". Therefore, transcriptions and other types of primary texts such as from books are not seen as annotations. Another usage (as is typical now in the field of language resources) creates too many irritations between the different disciplines although the definitions are more straightforward.



The two drawings indicate different dependency relations that normally occur when transcribing dependent channels (voicing can only occur while speaking, or hand shape change is part of a gesture).



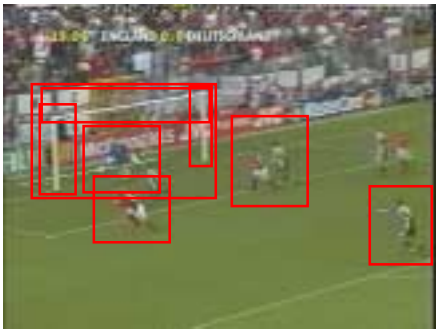
This example shows a typical example of a transcription or other basic text (yellow) that can be broken up for example into words (here drawn as separate tier). Associated with the words can be various other types of tiers. Here linguistic layers are indicated, but it could also be a comment on specific element in the text by person A that is commented by person B etc. The result is a hierarchical type of annotation where annotations at higher level refer to elements of annotations at a lower level.



These two example show the different forms of cross-references that occur in language resources. Cross-referencing within one tier and different tiers (of different files as well) are the most frequent ones. They can be uni- or bidirectional and can refer not only to structural units, but also to elements within structural units such as syllables although only utterances or words may be tagged explicitly. A complication is given if an annotation refers to one or several such units (again: can be in different files) and if the annotation containing the links itself can be subject of annotations.

2.2 Spatial Transcriptions and Annotations

Spatial annotations are very similar to time/sequence annotations except that the mechanisms operate in a 2D space. For matters of simplicity we ignore the time dimension as in videos, since 2.1 comments hold.



All kinds of spatial relations can be seen when annotating a spatial object such as an image. An annotation can be associated with any kind of polygon (does not have to be a rectangular shape). And an annotation itself can be subject of annotation. Similar as in the time domain we have “transcriptions” describing what is to be seen and higher layers of annotations that refer to the transcriptions or to annotations.

The same holds when we have a facsimile of a manuscript page containing various related objects such as formulas, texts, texts superposed to other texts (like corrections etc), drawings and photos. In the

same way the created transcriptions or annotations either refer to spatial objects identified by some polygon or to texts.

2.3 General Remarks

So no principal difference can be seen between time, sequence or spatial annotations. This has already been stated correctly in the ATLAS project [1]. A flexible XML-schema such as EAF [2] would have to be extended as suggested by ATLAS to cover spatial transcriptions and annotations as well.

A principal problem is given by the fact that often elements of structural units are subject of referencing. Assume that you have chosen to tag utterances/sentences as the basic unit as is depicted in the following example[2]:

```
<sent sid=”111”> My name is nobody </sent>
```

If you want to hookup a comment to the word “nobody” there is no way by XML technologies. One can use XLink to point to the structural unit “sentence with the id=111” but one has to mark somewhere that the 4th word is meant. This creates a number of well-known problems: (1) One could define delimiters such as a “space” to count words. However, it is known that for some texts delimiters are not systematically used. (2) One could adapt the sentence to a new version, since someone found out that the transcription was not completely correct. A new word would be added and all referencing based on word count would fail, i.e. existing references would become wrong. (3) One could just count the bytes at which the word (or unit) starts[3]. The same remark as for 2 holds, i.e. changes to the string will make existing links incorrect.

A solution would be in many cases that not the sentence is tagged as basic structural unit, but the words, which would deliver the following example (Dutch Spoken Corpus uses this schema):

```
<sent sid=111>
        <word wid=”111.1”> My </word>
        <word wid=”111.2”> name </word>
        <word wid=”111.3”> is </word>
        <word wid=”111.4”> nobody </word>
</sent>
```

Probably only in linguistic studies where one wants for example associate a tone with a certain sonorant cluster more fine grained referencing is required. Therefore, it may be worth to use “words” as basic structural units in ECHO since they convey meaning, i.e. semantic references will be associated with words or word groups. Other users needing more fine-grained access can help themselves with counting bytes within words assuming that the original text will never be changed.

2.4 Standoff Annotation

In the domain of XML files the standoff model is quite popular. It basically states that all different sorts of information are kept separately for easy management. This is certainly true for other type of bundling/grouping and references, since including this kind of information in the text is either impossible or would mess up the text with information that does not belong to it. Therefore, in XML references always should be kept separately and be treated as shown in the last example under 2.1. An example for a new grouping (into syntactical parts as an example) is shown in the following box:

```
<verbphrase vpid=”111.2” href=”word.xml#wid(111.3)wid(111.4)”/>
```

For cross-references of all sorts it would make sense to use RDF and immediately formalize the type of the relation.

[2] For matters of simplicity all discussions about where punctuation marks belong to etc are ignored here.

[3] This method was used in the Tipster model that was also applied within the GATE architecture.

It has to be checked carefully how far one wants to go with separation. In EAF it is possible to integrate many tiers into one file. However, they are separated within the file. However, the mechanism is flexible enough to allow storing tiers in different files. This is important with respect to collaborative environments and allowing random users to comment on texts. These different types of information should be stored in different files due to completely other reasons such as missing quality control and access rights reasons.

2.5 Schemas for Text Files

It is essential for ECHO to define a limited number of schemas for the various types of text files assuming that it is agreed that all files in ECHO should be available as XML files[4]. Every data provider has to make sure that his data is described by schemas. There are a number of requirements to be fulfilled for such schemas:

- The referencing to images, image parts, sound or video fragments has to be supported, which requires a specification in time and spatial units. A unique method should be defined. The EAF format does this by defining time points and periods that can be associated with time points. Periods can be associated with an unlimited number of annotations belonging to an unlimited number of tiers.
- The association of arbitrary annotations to textual units has to be possible by specifying the structural unit and some offset. EAF does this by introducing a sequence-reference construct.
- The schemas must support to describe the type of the annotations (its name, its language, its type-dependencies, etc). Comments are seen simply as special tiers. Footnotes etc can also be seen as special type of tiers where a footnote is an annotation associated with a structure unit and an element.
- For lexica a different schema has to be defined, since it is another text type with its own specific structure. It has to be seen in how far it is possible to come to "generic" schema for lexica that can cover all possible concrete lexica[5]. There may be other important text types requiring a different schema.

The ECHO standards should be in full accordance with established international standards where possible. New schemas should be defined only where absolutely necessary. We cannot accept that for every text file a new schema is created.

2.6 File Naming and Identification

We already have identified various different types of XML files such as

- metadata files (according to the IMDI and MIDAS schemas)
- controlled vocabularies for metadata or annotations
- transcription / annotation files for images, original texts, sound and video (according to EAF, ATLAS or other schemas)
- perhaps we need an additional file type to cover the complex texts one comes across especially in history of science
- files to cover cross-references of all sorts and that itself can be extended by annotations
- files covering lexica (and perhaps even other data types)
- others

We should make a complete list of different data types and choose common names within the ECHO framework. To a certain extent the naming conventions should express close relationships.

MPI Berlin has already published a paper discussing unique file identifiers. This is a topic to be discussed in more detail, since all references that will be established and contained in cross-reference files for example have to be stable. Metadata is used at the MPI Nijmegen for this purpose, but may be not sufficient.

[4] It would be acceptable that an XML file can be generated as output of a relational database as discussed with the EHESS team - independent whether it is created on the fly or permanently.

[5] In February a small group of international experts will meet to discuss the issue of a "generic" schema for lexica.

2.7 Tag Labeling

Interoperability is an important issue. This can only be achieved when it is stated explicitly at some place what the semantic relations are between the different structures within the ECHO domain. This holds for metadata as well as for transcriptions/annotations. Each data provider has to provide definitions of the type of labels he is using in texts. When a joint search has to be executed on resources we must assure that the tag labels “w” and “word” or “morpho” and “morphosyntax” may mean the same. Where necessary mapping files have to be created based on the provided definitions. These mappings should be written in RDF and made available as services on the web.

2.8 Visualization and Operation

In the ECHO discussions until now it became clear that the wishes for clever visualizations might be very dependent on the data types and the disciplines. Therefore, in ECHO1 it seems to be reasonable to continue with the discipline-based approaches. Similar aspects hold for operations on the data - be they semi- or fully automatic. A number of tools are under development and they serve the short-term needs of their community (more or less). It is up to ECHO2 and the work of the Technical Committee to discuss ideas of how to come with user-customizable and flexible environments.