Citation of Electronic Resources

- Proposal for a new work item in ISO TC37/SC4 -

D. Broeder, Th. Declerck, M. Kemps-Snijders, H. Keibel, M. Kupietz, L. Lemnitzer, A. Witt, P. Wittenburg[1]

This note is meant as a short presentation of the essential points to raise awareness and to stimulate discussions within ISO TC37/SC4. We leave the creation of a more elaborate paper to the phase after the Provo meeting.

1. Problem

One of our prominent goals is to intensify the linking between language resources of different types and across languages. This strategy will enhance, for example, the semantic exploitation of these resources to the extent that such linked domains offer ways to access resources via semantic ports. In addition, publications will increasingly often refer to electronic resources or fragments of them. This trend toward interlinked resources will not work if we do not have a mechanism for uniquely and persistently referencing these resources. Traditionally researchers have tended to refer to language resources by their proper names (e.g. Brown Corpus). This practice is no longer adequate: The available resources are constantly growing in number, and many of the new resources are dynamic in nature, such that it may not always be clear what exactly such proper name would refer to. The current practice of using URLs is not satisfying either, since it mixes “referencing” and "physical storage", although we all know that physical paths are continuously changing for example due to technological innovation. This whole issue was discussed at the last DELAMAN[2] meeting that was held in November 2006 in London and received a high priority.

The issue can also be compared with ISBN numbers in the domain of physical publications (books, CDROMs, etc) where a registration office ensures that every object (not its many serial instances) receives a unique and everlasting "string" representing the object. However, in traditional citations the ISBN number has not generally been used, since it cannot be directly interpreted to allow meaningful inferences. In our Internet era this aspect, however, has changed: everyone can activate a link and start appropriate services, i.e. the resolution of the string does not require human actions and could, for example, lead to a metadata description that contains all the details about the resource or resource collection. There is no need anymore to include typical reference information such as author, title, etc., in the immediate reference for a direct interpretation because this information can be retrieved or generated if needed. However, sometimes printed versions may be used in parallel.

2. Type of Resources

Although we believe that the proposed solution will work for references to electronic resources from almost all disciplines, we will restrict ourselves to the language resource domain. Here we can distinguish between the following resource types:

- single atomic resources
- bundles of related resources
- collections in the sense of published corpora
- collections in the sense of incidental groupings of resources that were used in a particular scientific project and that are derived from other corpora

Atomic resources include all linguistic resource types that we know of, such as text documents, annotations, audio/video files, time series such as those generated by laryngographs or eye trackers, lexica, grammars, metadata descriptions and ontologies. Bundles can be of various sorts such as annotated sound files, lexica with media extensions, etc. It must be possible to refer to such a bundle of closely related resources. With respect to collections it makes sense to distinguish specific corpora

[1] The authors are partly members of the German and Dutch ISO groups.

[2] Digital Endangered Languages and Music Archive Network, a group of experts coming from many of the most relevant archives world-wide storing resources from endangered languages and cultures. Two papers were given: one by Jeff Good (MPI Leipzig) and one by Daan Broeder (MPI Nijmegen).

that are created by a project and have a certain status, such as the Brown Corpus, the National British Corpus, the Dutch Spoken Corpus, etc. These corpora normally have an elaborated metadata description that refers to the collection as a whole. It must be possible to refer to such a corpus as a single unique resource.

In addition, we have virtual collections[3] that are created temporarily by individuals or projects to support a certain research work by combining parts from different other corpora. Such collections mostly do not have a value as a whole except for the specific project at hand, and they will normally be collections of resources collected from various repositories. In the case of collections that support a dissertation that compares phenomena in several languages, for instance, the number of included resources may become extensive, i.e., also in this case it makes sense to simply refer to one collection. This could be done by asking the researcher to create a simple metadata description that links to the different contributions. It would not even require integrating all the information in this one metadata file, since this could be an enormous effort. It would be sufficient to allow the interested user to navigate in the enclosed set or hierarchy of metadata descriptions to find out what it contains. Such virtual corpora do generally not have the status of published corpora, but it should be possible to reference them as a whole. This kind of documentation necessary because results of empirical research based on such virtual corpora must be traceable.

In addition, it is necessary to be able to refer to a fragment of a resource. Here we need to differentiate between the different linguistic data types:

- in a structured text we probably want to refer to a structure element (X-Path), in a lexicon for example this could be the whole lexical instance, a lexical entry or a lexical attribute in a lexical entry
- in a text document or within a text element we probably will use a character offsets
- in a sound file we will use start and end time
- in an image we will mostly use two coordinate pairs (in case of rectangular marking shapes)
- in a video we will use both coordinate pairs and start and end times
- in time series data we will use the channel number and start and end time
- a relation in an ontology will be referenced probably by a unique identifier

3. Unique and Persistent Identifiers (UPID)

The core of a reference in electronic resources are the unique and persistent identifiers (UPID or DOI) that are provided to refer to a collection or bundle, a resource or a fragment of a resource and that can be resolved to real resources or resource descriptions. Additional information can be given immediately where it makes sense, but as stated above it may all be described in metadata descriptions that can immediately be invoked.

There were different suggestions to overcome the limitations of using URLs as a reference mechanism. PURLs pretend to be persistent URLs, however, they still mix location and protocol issues and the resolving mechanism depends on the HHTP protocol and has a single point of failure. More widely agreed in the Digital Library world is the Handle System which provides UPID/DOI specifications and also a resolving system. The syntax of a handle in the Handle System is very simple:

<prefix>.<suffix> ***example: 15.12345abcd6789***

The Handle System top authority will assign the prefix on request to an institution or organization and will be able to resolve any such Handle, i.e. by its Global Handle System it will be able to identify the prefixes, the prefixes will point to Local Handle Systems and these will know how to interpret the suffixes, i.e. the specific syntax of the suffixes is left to the Local Handle System owner as long as it meets the URI specifications. In the following we give a few examples taken from the DOI web-pages:

10.1000/123456, ***10.1000/ISBN1-900512-44-0,***
10.2345/S1384107697000225,
10.4567/0361-9230(1997)42:<OaEoSR>2.0.TX;2-B,

[3] Examples are DEREKO (the Mannheim German Reference Corpus) accessed via COSMAS II or the IMDI Infrastructure from MPI.

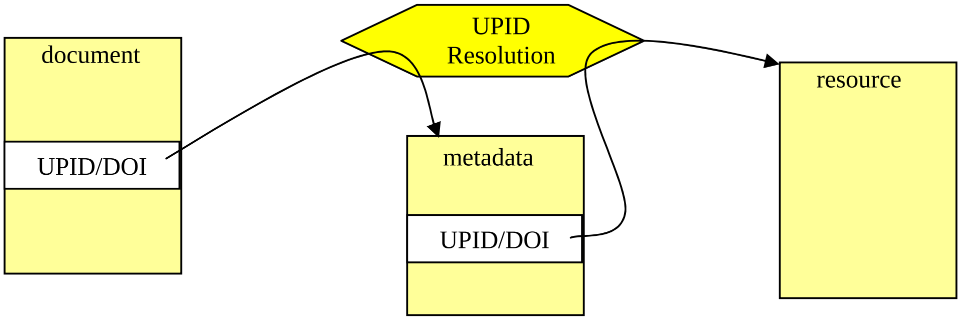
10.6789/JoesPaper56

General resolving scheme:
1. prefix resolved by Global Handle System to Local Handle System
2. suffix resolved by Local Handle Systems to URLs

Such handles conform to the functional requirements of the two generic approaches for naming first-class objects on the Internet: the Uniform Resource Name (URN) and the Uniform Resource Identifier (URI), and handles need to be used with other http-based mechanisms such as OpenURL, PURL, parameter passing etc. Another benefit of the Handle System is that a single handle can be resolved to multiple URLs thus allowing usrs to identify copies of a resource with the same handle. This document does not intend to discuss all details of the Handle System (see http://www.**handle**.net). Prefixes can be requested from the Handle System group.

The DOI Foundation architecture is based on the Handle System and acts as a handle authority and introduces another layer. Although DOI object identifiers specify their own URI protocol, they contain a handle that has the prefix "10" such as *DOI:10.1007/s003390201377.* As can be seen in the example DOIs, the suffix has two parts: again another kind of sub-prefix that is specified by the DOI registration authority and issued to local repositories. Thus, the principles remain the same and it must be left to the initiative whether it will collaborate with the Handle System group of with the DOI Foundation. Obviously the Global Handle System will also be able to point to the DOI resolution system when an attempt is made to resolve the reference from another handle prefix domain.

It is suggested that ISO TC37/SC4 make a statement asserting that the syntax of the Handle System will be adopted as the basis for the referencing mechanism. This handle will be resolved (1) into one or several URLs that point to instances of an object, object bundle or collection, or (2) into one or several URLs that point to instances of metadata descriptions of an object, object bundle or collection which will include UPID/DOIs that point to the resources. Clarifying this issue for our community will make it obvious that we all will create an interpretable and interoperable domain.



Representatives from TC 37 will need to discuss with the Handle System Group of the DOI Foundation how we can ensure redundancy, high performance, independence, availability, persistence.

5. Addressing Fragments

As indicated we need mechanisms for addressing fragments in resources that are identified by UPID/DOIs.

Here we suggest that ISO TC37/SC4 create specifications for the resource types that we have identified so far. These can be expanded stepwise. For a sound file we would have something like in the following example:

1839/00-0000-0001-4C55-3#time(ms):23680,24759

The fragment identifier indicates the fragment type as mentioned (in this case "time"), the unit of reference (in this case "ms"), and the start and end times. These specifications have to follow guidelines that are specified within other standards, for example to refer to video timing and frames or to use X-Path to denote a fragment in an XML-based resource.

4. Additional Information

We have to distinguish between different referencing contexts. In some contexts, such as an image in a lexicon, the user is not interested in seeing some information about the image in addition to what he has already found in the lexical attributes. He will only be interested in activating the reference and seeing the image itself immediately. In other contexts that are meant for a human reading a reference, the traditional style might be more appropriate. In the first case the reference is just the UPID/DOI with a fragment indication, if necessary. In the second case we need to have a closer look.

The UPID/DOI is not very informative, as already indicated, and the credits to the creators are not immediately visible. The DOI Foundation gives a few examples how references could look, including both types of information: the traditional type of information and the UPID/DOI. The first example is more of a typical reference to a publication where the UPID/DOI is added. The second is to a published article.

- Kornack, D. Rakic, P. (2001). Cell Proliferation Without Neurogenesis in Adult Primate Neocortex. Science. 294 (5549), 2127-2130, doi:10.1126/science.1065467.
- "Cell Biology: A cat cloned by nuclear transplantation" Nature AOP, Published Online: 14 February 2002, doi:10.1038/nature723.

ISO TC37/SC4 does not have to make recommendations or specifications about what kind of information needs to be integrated when we are referring to traditional publications. There are various quasi-standards defined by publishers and research organizations. We should require, however, that resolvable UPID/DOIs are added in case of web-accessible resources.

In case of resources, bundles or collections, there is no clear suggestion yet, since each individual resource can already have many authors. This gets even more complicated in the cases of bundles or collections. We only can supply a template stating that (1) depositors/authors/collectors/creators, (2) the year of publication, (3) a title/name and (4) the archive where it is stored could be cited.

The metadata description is assumed to include all this information so that interested persons can look it up.

The specification of version information is not a matter to be handled here. Every new version of a resource needs to get assigned a new unique identifier, and it is left to the Local Handle System whether and how to explicitly specify the version in this identifier. Version information can be included in the metadata description. For the case of collections it is not at all obvious how this should be handled, since in the case of many resources that are subject to frequent changes, the number of versions at the collection level may rise almost infinitely. Here different archives use different strategies; one could be to associate time stamps.

5. Rendering

Any matters that have to do with visualization of certain aspects are not part of the ISO TC37/SC4 recommendations. The UPID/DOIs will appear as references in very different contexts. It is the task of the appropriate viewer to present the document including its references in a suitable form and it is the task of the metadata infrastructure to present the metadata in an appropriate form.