

Figure 1. Links referring to physical addresses

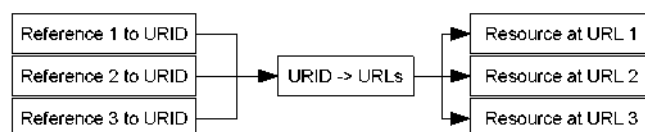


Figure 2. Links referring to URIDs

Unique Resource Identifiers

Daan Broeder, Eric Auer, Peter Wittenburg
MPI, Nijmegen

Many valuable language resources have become available via the Internet. These include primary resources such as media recordings, secondary resources such as annotations, lexica and grammars, and other electronic publications. Increasingly, there are also references and links between resources: entries in lexica refer to fragments in annotated media recordings, and grammars refer to examples in primary sources.

Currently, the typical way of making such links is via hyperlinks based on URLs (Uniform Resource Locators). However, URLs are unstable identifiers, compared, for example, to bibliographic references used for books and journals. URLs can change as servers and storage systems are replaced or reorganised by system managers, or as resources are moved between institutions. And when such changes do occur, it is practically impossible to modify all existing references to URLs. Therefore, URLs do not provide a stable and persistent method of referencing.

Unique Resource Identifiers (URIDs)

We need a mechanism that can overcome this instability. Valid references need to be maintained while the locations of the resources change. A solution is to

introduce a global naming system that uniquely and persistently identifies each electronic resource separately from its URL address. We call the identifiers “Unique Resource Identifiers” or URIDs. These can be compared to ISBN numbers; they uniquely identify titles and provide a distinction between the information object and its physical instances at particular addresses.

Like ISBNs, URIDs are administered by institutional registries. These registries provide the service of translating URIDs into actual web addresses, so that the resources can be accessed. This mechanism can again be compared to the world of books: one looks up a book’s title but then has to get an “address” (i.e. the book’s library call number) in order to locate it on the shelves before reading it. In other words, using URIDs means introducing the complication of another layer of reference between resource identities and their instances. However, the advantage is that when resources move, the addresses only have to be administered in one location – the URID registry.

In the existing system of web links, several links may point to a given resource at its physical address (Figure 1). By contrast, in a URID system, all links refer to a unique and persistent identifier, which in turn is mapped to one or more URLs (Figure 2). Incidentally, the figure illustrates the case where there are multiple instances of the same information resource, for example in different archives.

The main challenges in introducing this additional layer into the linking mechanism are that the mappings of URIDs to URLs have to be done with care, and that the online services that provide the mappings have to be available 100% of the time.

DAM-LR and the Handle system

In the DAM-LR project (Distributed Access Management for Language Resources), the participant archives chose an architecture that combines the advantages of URIDs with the flexibility to cater for the particularities of each archive. We chose the Handle System, already widely accepted in the world of digital libraries and archives, as our URID (or “handle”) registry and resolving (mapping) system. A complete URID (handle) is a combination of a centrally-assigned prefix and a

locally-defined postfix. The Handle System registers institutions as local authorities with their own URID prefix. The institutions can then locally specify the remainder of the identifiers (the “postfixes”), map the identifiers to URL paths, and provide a public service mapping handles to paths. The global Handle System authority directs references that cross prefix domains to the appropriate local mapping service.

Next steps

Participants in the DAM-LR project registered as authorities at the Global Handle Registry service. This allows them to maintain their own mapping services and operate independently, while still enjoying the benefits of URIDs.

URIDs as described above refer to complete resources. We would ultimately also like to be able to refer to relevant fragments within resources, such as segments of video or entries in a lexicon. For XML-encoded text documents, handles could be combined with XPointer references to achieve this; however, a general mechanism is yet to be defined. Nevertheless, URIDs provide the first step in solving the question of how to create a reliable system for referring to electronic documents.

Links

DAM-LR: <http://www.mpi.nl/dam-lr>

Handle: <http://www.handle.net/>