Citing on-line Language Resources

Daan Broeder, Dieter van Uytvanck, Gunter Senft
Max-Planck Institute for Psycholinguistics, Nijmegen, Netherlands
E-mail: {daan.broeder, dieter.van.uytvanck gunter.senft}@mpi.nl

Abstract

Although the possibility of referring or citing on-line data from publications is seen at least theoretically as an important means to provide immediate testable proof or simple illustration of a line of reasoning, the practice has not been wide-spread yet and no extensive experience has been gained about the possibilities and problems of referring to raw data-sets. This paper makes a case to investigate the possibility and need of persistent data visualization services that facilitate the inspection and evaluation of the cited data.

Keywords: citation, visualization, persistent identifiers

1. Introduction

Citing Language Resources is essential when an author creates (electronic) documents and wants to illustrate some reasoning, provide some easily accessible examples or even provide an immediate testable proof for a theory. The classic way of citing resources and documents from papers has been standardized and described by many recommended practice documents [1], [2]. However, the possibilities of citing on-line available resources and documents are not well know and have not been fully exploited as for the majority of existing resources this kind of instant availability does not exist – with only some more recent exceptions.

We differentiate between documents and resources because the citing information for documents has been far better standardized. It is also straightforward to present a suitable view of the cited document thanks to the many eDocument servers that are available. In the remainder of this paper, however, we will only speak of Language Resources implying that documents are just a sub-type of these resources

The citation of a Language Resource in an embedding document can be viewed as the combination of the human readable citing information and a (preferably) actionable identifier that can be used to call up an adequate presentation of the resource. As mentioned above, the citing information for documents is already well standardized and formalized and for slightly more exotic formats, like media files, standards exist as well. However due to the large variety of resource types, no generally accepted standards exist and conceivably only guidelines can be given [2], [3] that target aspects like proper acknowledgement and adequate description.

For the purpose of this paper the more interesting part of the citation is the string that identifies to the resource itself and can be used to locate, access or visualize the resource. Ideally this identifier is actionable, that is to say that when viewing the document with a suitable viewer or web browser, the user can perform an action on this identifier. This is usually achieved by having the resource identifier encoded as a URI string, which will become automatically actionable when displayed in a web browser and most document viewers. The desired result of this action is that the resource is presented to the user in a usable form. A minimal option would be that the user can save the resource on his computer, but this is clearly suboptimal for fast inspection. Having the resource immediately displayed in some useful form is a better option in most circumstances. This can for example be achieved by having locally installed viewer applications on the user's computer where data-types are mapped to specific viewers based on their mime-type. This is of course standard practice for common data types and can be extended to specific linguistic data types such as the ELAN multimedia annotator for mime type "*text*/x-*eaf+xml*".

Another option is having an identifier in the citation not directly refer to the resources but rather to a resource display service web application that is able to adequately visualize it. If the citation is embedded in a web document, the resource can be displayed in-line, much in the same way like for example a YouTube video on a web site. The option of using a web application for visualization is considered to be most versatile because it allows the development of highly specific web applications (viewers) specially created for this purpose. This approach does not require the installation of special viewer applications on the user's computer.

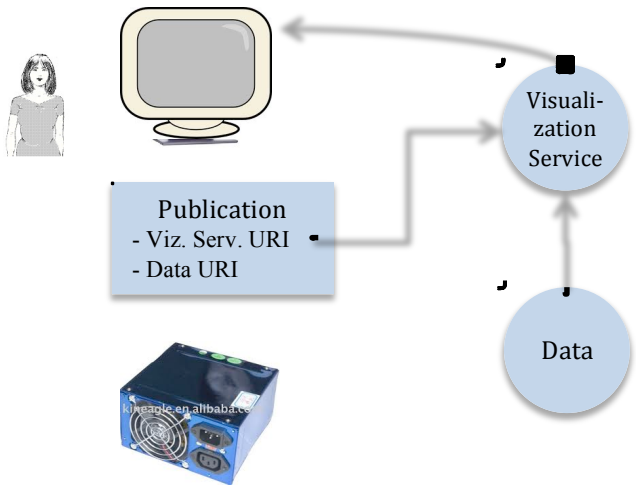

Figure 1 Citing data from a publication and including a link to a visualization service thus allowing visualizing the data next to downloading the data.

2. Issues

Considering the previous discussion the following issues come to mind:

1 Persistency of the resource. Usually the identifier is a URI, the stability of which is constantly debated. Solutions in the form of Persistent Identifier Frameworks (PID) have been created. A useful paper explaining these issues especially for the LR domain is [3]. Currently the dom

2 When using a PID framework such as the Handle System (HS) [4], it is possible to provide both the resource location and the URI of a visualization service with the resource's PID.

3 Persistency of the display service. The service is responsible for displaying the cited resource in the desired manner, usually by means of a web application. This service is started with the resource identifier as a parameter. For the purpose of resolving the citation, the display service must be as persistent as the resource identifier. It is not necessary that the same provider as the resource hosts the display service as long as the viewer also allows downloading the resource. If the display service does not offer a separate download option a separate download link should be part of the citation. Downloading the resource allows the user to process the resource locally. At the TLA-MPI archive we offer such services for EAF type annotation files. One for embedded display in web-pages and one as a separate web-application ANNEX [5] [6].

4 Service versions. Having new versions of the display service should be acceptable as long as it performs the same type of visualization. This is in contrast to the resource where the identifier is (in by far the most cases) required to point to the original resource.

5 Data granularity. Citation should also be possible for parts of resources like a fragment of a text or a media file or an entry in a lexicon. This requires that the used identifier can hold information specifying the part intended. See [3] for an explanation of part identifiers.

6 Data versioning. It can be useful to show the development of the annotation process by showing different versions of an annotation file. A repository that stores these different versions can support a visualization tool that allows a user to step through subsequent versions.

7 Display format. The way in which a resource should be visualized optimally is dependent on the resource type, on what the citation creator wants to achieve with it and what the end-user requires (e.g. maximal resolution of a picture vs. a thumbnail). That last point can be handled by mechanisms as HTTP content-negotiation.

8 Citation creation. Usually archive catalogues or publishing platforms offer the citation information together with that string for copy and paste actions in the GUI.

9 When discussing the use of referencing visualization services from publications, many researchers have expressed reluctance to rely on possibly inherently 'unsustainable' or 'unreliable' service. Compared to the familiar referencing of stable static data records this reluctance seems justified, irrespective of the efforts taken to make such services stable. Therefore it is important that the citation also includes a reference to the data record itself, not only to allow direct downloading of the data but also to reassure the researcher.

.

Resource Type		**Display method**		**Citation context**
Resource Type		**Display Method**		**Citation Context**
Metadata record		In-line		Example
Primary text		Separate web application		Illustration
Annotations		Download & local viewer & tools		Evidence
Lexicon				Counterexamples
Audio file				Additional reference
Video file				*Combinations of the above*
Image file				
Audio analysis (e.g. pitch)				
Relations between resources. E.g. RDF				
Combinations of the above				

Table 1 Three dimensions for LR data citation

Although much can be said about all these issues, in what follows we prefer to elaborate further on the "Display format" , "Service Persistency" and "Citation creation". The other points have existing recommendations [3] or are more or less self-evident.

3. Display Format

The display format is a subject that can only be approached in a heuristic manner and is usually also dependent on the purpose of the citation.

We discussed possible options with researchers from the MPI for Psycholinguistics the DoBeS [7] project and others to get their view on existing facilities and the need for new ones. We will continue discussing and analyzing these issues with researchers and have put this on the agenda of a larger communities as CLARIN [8], [9]. These discussions were based on three categories: "Resource Type", "Display Method" and "Citation Context". Table 1 shows these three dimensions with their respective value ranges.

With respect to the resource types, it was stated that although in some cases a simple media player (sound only or video) is sufficient, it is often necessary that the video file be accompanied by additional material, such as time-aligned morpheme-glosses or free translation.

In linguistic research, for example, a serious analysis of phenomena like serial- or multi-verb constructions requires to have access to the physical sound signal in the transcription to decide, illustrate and prove that such a construction really constitutes one intonational unit.
Another example where the combination between traditional text and access to original data has proven to be extremely useful is an anthropological-linguistic project carried out at the MPI for Psycholinguistics in which an author documents and analyses various genres that are differentiated by the researched speech community. This verbal documentation is linked to the actual data and thus allows checking the presented data transcription with the original speech data. With ritual texts like specific songs etc. the combination of the printed analyses with audio- and video documents open up a broad framework with respect to the complex semiotic systems in which these text genres are embedded [9]. It is a desideratum for many researchers (not only) in the social sciences to link texts that are based on transcripts of interviews or other data gathered in participant observation to the original primary data that are not completely or even not at all presented in the respective scientific paper.

Although most of the different values for the citation context dimension do not require different functionality, the "Counterexamples" suggest the need for a facility to display at least two resources simultaneously. Also the "Illustration" and "Example" might be sufficiently served with a limited size in-line display while for a "Evidence" context visualization with a reliable tool such as for example an audio signal viewer with analysis functions is required.

Research on the conceptualization of space and the various frames used to refer to objects in space carried out at the MPI for Psycholinguistics revealed that there are basically three different frames of spatial reference in the languages of the world. Data were collected using identical methods for data elicitation. To illustrate the functional equivalence of these frames of spatial reference a "Counterexamples" device would have been extremely useful.

In all cases where more than simple illustration is intended, the availability of download function next to visualizing the resource was thought necessary.

4. Persistency of visualization & display services

It was already pointed out that the display service could be a service that is separate from the service delivering the resource. Furthermore, these two services can be run by two different organizations. This suggests that it is advantageous to create a few well-equipped powerful display service sites that can be used by the LR community to display cited resources from all over the community. Clearly, we refer not just to the visualization of simple data-types, but also to the visualization of complex structured data such as metadata, annotations and lexica. For example some instances that come to mind are viewers for IMDI, OLAC and CMDI metadata and EAF and TEI annotations.

Prerequisites for such a separated setup is that (a) the resource formats are (de-facto) standardized by the user community (b) the authentication and authorization problems are solved and (c) the offered display formats are general enough to satisfy the larger part of the community.

Having both the resource server and display server use the same federated login and "single sign on" mechanism as proposed by amongst others, infrastructure projects as CLARIN will solve the authentication and authorization problem. The acceptable "general" display formats is a matter of careful consideration of community requirements. Here it is as well on this issue on which the present paper hopes to start a discussion.

Organizational issues in connection with the service persistency requirement are possibly the most challenging task. Thanks to resource infrastructures like CLARIN, the LR community is gaining familiarity with the resource persistency issues and the respective solutions: PID frameworks, stable cooperating repository systems etc. However, issues around the persistency of services, are relatively unexplored. In the Dutch CLARIN project a number of so-called service demonstrator projects were started where researchers were asked to create a LT service that should run persistently at a CLARIN center. Early results from this indicate that persistency of services is far more difficult than handling persistency of data and requires CLARIN centers to impose special requirements on such services. It is clear that the introduction of display services bound to citations have as absolute need to be persistent over

time, will push this work forward.

5. Citation creation

It should be one of the functions of language resource catalogues to provide adequate citation strings that can be copied and pasted into papers. Likely multiple formats should be produced on request, so that users may for example choose between APA and ISO 690 conventions. It should be stressed that these citation formats are used in combination with a direct pointer to the resource. In the case we refer to a persistent and reliable visualization service, this service can also be used to display any citation information, making perhaps another citation format more suitable.

If a publication requires the citation of very many data resources it should be considered to create a Virtual Collection (VC). Such a collection is a metadata description that describes and links to many different (possibly distributed) resources. It differs from normal collections in that it was created specially for the research purpose or publication. Of course such a VC must be given persistent existence in a virtual collection registry (VCR) just as the resources themselves, with the same life expectancy as the constituent resources of the collection themselves. VCs and VCRs have been the subject of discussions within CLARIN [11]. The Virtual Collection Registry should of course, just as an archive catalogue, produce citation strings in appropriate formats. In the case that we want to have specialized visualization services also to work for such a VC, they have to be able to recognize such collection descriptions and allow the user to first browse through the constituent resources and on request visualize them. Having a special VC visualizer and storing the PIDs for the respective resource visualization services in the collection description next to the PIDs for the actual constituent resources can realize this.

When creating a citation, this is one of the two moments in the life cycle of the data object that a user actually comes into contact with the PID of an object. In our case it is part of a bigger formatted text object and unless he pays special attention, it is transparent to the user. This is in contrast with the second encounter that occurs when a user actually needs to access the cited data and he consciously has to click on the actionable PID in the citation text. If the PID is not a URI, it is not automatically actionable. In that case it is recommended to provide the PID in an urlified form so it becomes actionable. There are plug-ins that will make www-browsers and document readers handle non-URI type of PIDs actionable. However it can not be relied upon that every browser and document reader has been equipped with such a plug-in.

6. Summary and perspective

We propose to investigate the need and possibility to mount a number of persistent visualization services for complex LRs and use citations that directly link to these services. This does not replace the need to reference to the LR itself also; indeed I t is imperative that a resource down load is possible also, either via information associated with the resource's PID, the display service itself or by adding an extra reference in the citation.

This effort of investigating the use of visualization services in citations is only part of the effort to come to an accepted and standardized format for citing LRs. Important for this are the ISO efforts [2] wrt. the citation text and [3] wrt. persistent identifiers. It remains to be investigated if these are sufficient to meet the requirements of the LR community. The community represented ao. in FlareNet is already aware of some of these issues [12] but perhaps more action is called for as was suggested by some.

A possible course of action can be to investigate if the current available standards and practice of APA and ISO 690 citation texts are sufficient for the LR community especially when using references to visualization services as is proposed in this paper. This is a matter of careful consideration, since the LR community should not deviate from the practices in other disciplines without good reason, and inconsiderate deviation might also break important tool interoperability. Therefore also with respect to such important general data management services as PID frameworks the LR community should not experiment but rely on larger proven systems and initiatives as for example EPIC [13] and DataCite. [14].

7. References

[1] APA. Style Guide to Electronic References, June 2007, ISBN: 1-4338-0309-7

[2] ISO 690:2010. Information and documentation – Guidelines for bibliographic references and citation to information resources

[3] ISO 24619 PISA Language resource management — Persistent identification and access in language technology applications, http://www.clarin.eu/files/iso_tc37_sc4_N626_wg1_DIS_24619_PID.pdf

[4] Handle System, http://www.handle.net/

[5] ANNEX manual, http://www.lat-mpi.eu/tools/annex/manual/

[6] latnews June 2010, http://www.lat-mpi.eu/latnews/2010/06/embeddable-annex/

[7] DoBeS http://www.mpi.nl/DOBES

[8] CLARIN, http://www.clarin.eu/

[9] Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. *LREC 2008*

[10] Senft, G. (2010). The Trobriand Islanders' ways of speaking. Berlin: De Gruyter

[11] http://www.clarin.eu/files/virtual_collections-CLARIN-ShortGuide.pdf

[12] Calzolari, N., Ouochi, V., Soria, C. () FLaReNet Strategic Language Resource Agenda, from http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf

[13] EPIC, http://pidconsortium.eu/

[14] DataCite, http://datacite.org/