Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

February 2010

Relevance

CLARIN community

for all communities

Virtual Collections

What is it?

A virtual collection is an aggregation of various data resources that serves a certain research purpose and that covers resources from various repositories most probably generated by different researchers and teams. There are various scientific motivations to create such virtual collections and give them an own identity, i.e. a separate metadata description and a persistent identifier so that it can be cited.

What is it for?

Users increasingly often want to

- access large samples of language data that are sufficiently representative with respect to the specific research domain and question they wish to investigate;
- preserve the context of their research which can be a collection of various resources and resource types; and will facilitate replications.
- create large virtual collections of similar data to identify parameters of stochastic engines such as Hidden Markov Models or Finite State Transducers;
- virtually bring together the few resources that are available for a minority language but distributed across archives;
- organize a (virtual) collection in a way that is suitable for their work process, i.e. combine metadata representations into self-defined views and hierarchies.

While off-the-shelf corpora cannot meet these needs, virtual collections defined on the basis of metadata and internal properties could. For all these activities users want to abstract from the idiosyncrasies of the repositories the resources are originating from. Of course, users would also like to see all interoperability (formats, semantics) problems solved so that they can work seamlessly on the virtual collection. Basis for the creation of virtual collections is a joint domain of compatible metadata descriptions as is being worked out in CLARIN. Thus, virtual collections can boost the re-usability of existing resources and facilitate empirically sound e-Science in the arts and humanities.

Who can use it?

- Since metadata is open and can be combined in a self-determined way any researcher can build virtual collections.
- Of course finally users also want to access the resources that are aggregated in such a virtual collection. This will require access permissions. The domain of trust CLARIN is looking forward to establish will simplify operation for those who are registered as users within an identity federation, since a single-sign-on would be sufficient. However, users also need to accept access terms and ask for access permissions in many cases.

When can it be used?

The possibility to create virtual collections from a first set of integrated centres should become available in 2010.

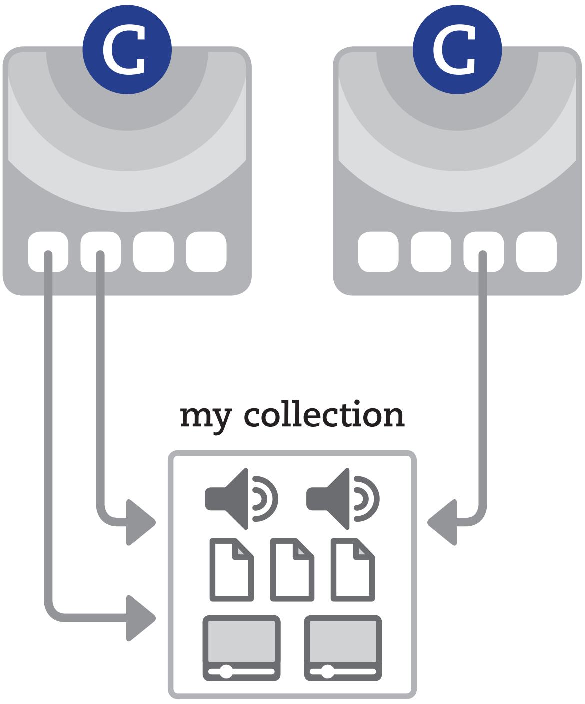
How does it work?

The following diagram can help to explain the principles which differ from the normal procedure that a user selects a resource by metadata browsing and searching and then opens it with the help of browsers or other applications as is indicated at the left side.

When building a virtual collection in general the user will select a set of useful resources based on metadata search and browsing. The diagram indicates that there is a joint metadata domain with a root description at the top and two example repositories joined under this root. The user will now select resources by virtually adding metadata descriptions into the basket which represents a new collection, i.e. only metadata descriptions are manipulated, selected, copied and brought together. The user can select new browsing hierarchies and views on the selected material which is optimally suited for his goals. Finally he will associate a metadata description and a PID with his new root node which may be called "mycollection". This root node will cover all references to the parts meaning that scientifically appropriate acknowledgements can be generated automatically. Very important is that the resources themselves don't have to be moved at all to build a virtual collection, metadata descriptions just point to them.

CLARIN will take care that users can create such virtual collections in workspaces offered as a service by infrastructure centres. Applications will then operate first on the metadata of "mycollection" and then contact the real resources if access permissions are given. Virtual collections can be dynamic. In that case snapshots and versioning are required to allow utilisation and replication.

It will be possible to define collections not only by explicit references to their parts, but also by "intentional definitions" such as "one million words corpus of randomly sampled newspapers of last week".



Who is responsible?

Within CLARIN work package 2 is responsible to offer this possibility. Two tracks need to be continued to make this possible: (1) The first federation of CLARIN centres needs to be established and linked with some identity federations. (2) The new component-based metadata infrastructure needs to be developed.

Whom to contact?

For all aspects the following address can be contacted:

Dieter van Uytvanck (MPI): dieter.vanuytvanck@mpi.nl

Where to find more information?

Most of the information about this feature is available via the information about the federation and the metadata framework:

CLARIN: http://www.clarin.eu/specification-documents

Responsible for the content:

Dieter van Uytvanck
MPI for Psycholinguistics
Wundtlaan 2, 6525 XD Nijmegen, NL
Website: www.clarin.eu
Email: dieter.vanuytvanck@mpi.nl

The last version of this document is maintained at the CLARIN Web-Site under documents: www.clarin.eu/documents