### A. The scope of CLARIN

CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities. CLARIN will serve the different linguistic communities, humanities scholars and the society to enable lower thresholds to multicultural and multilingual content. CLARIN is devoted to create a Pan European infrastructure that will offer persistent services that are secure and will provide easy access to language processing resources.

CLARIN is aware of the necessity to address the challenges of an increasing interest in electronic and multimedia communication. If we want to preserve cultural identity on the one hand, but make younger generations ready for the global competition on the other hand we have to improve their multilingual and multicultural awareness. CLARIN wants to make essential contributions.

### B. What CLARIN can do today in data-management

The CLARIN center federation consists of universities and research institutes throughout Europe. These centers provide an infrastructure for the distributed access and processing of written, spoken and multi-modal language data.

The data covering a wide range of human languages has a heterogeneous character. This includes various kinds of semi-structured data like text corpora and audio recordings, covering a broad extent in size and complexity. CLARIN offers centralized access and net-based search for these language resources and tools.

CLARIN has put in place Service Oriented Architectures, Single Sign On systems and metadata administration. The flexible way the infrastructure was implemented makes it easily adaptable for integrating it into other user communities.

### C. Why CLARIN needs EUDAT

The above described CLARIN center federation has two weak points which make scaling difficult with respect to a high number of simultaneous users and seamless data flow: lack of computing power and lack of efficient management of the data lifecycle.

Computing power:

* Accessing the data (e.g. large text corpora, in the range of millions of words). These corpora exist but cannot be accessed efficiently due to their semi-structured nature. More enhanced data accessing methods than simple full-text search are necessary.
* Extensive enriching and filtering the data (e.g. in the form of map-reduce) is performed, making high computing power necessary.
* Data exploration and exploitation of filtering results demands intensive computing power.

Managing the data lifecycle:

* During the data-lifecycle, data will change in type and increase in size and complexity. Flexible management of mass storage is needed.
* Documenting provenance needs versioning of the data.
* Archiving the publishable version of the data (PID, generic metadata, etc.)

EUDAT is needed to provide generic solutions to data access and storage.