CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

Data in CLARIAH

The increasing availability of massive quantities of digital data is one of the main reasons why an infrastructure project such as CLARIAH CORE is needed. The massive amounts of the data make it impossible to research them in the traditional way. The researcher has to use digital software to aid him/her in finding potentially relevant parts and ignoring irrelevant ones, or to carry out analysis of the data. But using software to search in and analyse massive amounts of digital data actually creates new opportunities for breakthroughs in humanities research, since it can be based on more data than ever before possible, and since it can make use of automatic analysis software that is more reliable in certain search and analysis tasks than humans are or ever can be (though in others humans still beat software).

Data come in many types. The major types are natural language texts, audio-visual data and structured data (databases). All three types are represented in CLARIAH. Though all types occur in all of CLARIAH's core disciplines, each core discipline has its own dominant data type:

- Linguistics: natural language texts
- Social economic history: structured (often quantitative) data
- Media Studies: audio-visual data

In addition, a discipline-independent work package deals with data that are useful or needed for all humanities disciplines.

WP3

In the linguistics work package (WP3), natural language texts play an important role. For some research questions the texts are sufficient as such, but in most cases the texts must be enriched with linguistic annotations such as part of speech tags for occurrences of words, full syntactic structures for occurrences of sentences (treebanks), and many other types of linguistic annotation. Searching in these linguistically enriched data requires special applications. Both the software to enrich the textual corpora (Frog, Alpino, Namescape, etc)[1] and applications for searching in the enriched corpora (OpenSONAR, PaQu, GrETEL, MIMORE, and others) were available before the start of CLARIAH-CORE or are being developed in independent projects (Nederlab) but many are extended and improved in CLARIAH. The data are large and distributed over multiple centres, so it is necessary to be able to search in such distributed data: search applications that can deal with such distributed or even federated search will be developed in CLARIAH. The major centres for natural language texts are Meertens Institute, Huygens Institute, Institute for the Dutch Language, and DANS.

WP4

In the social economic history work package (WP4) structured databases play a dominant role. Information on social economic history is encoded in databases. The relevant information concerns

[1] See http://portal.clarin.nl/ for more examples

several levels: the micro level (individuals and families), the meso level (organisations, trade unions, guilds, etc.) and the macro level (national and supranational data). The problem is that each database has its own structure and uses its own vocabulary. As a consequence, there is neither syntactic nor semantic interoperability. WP4 aims to address this problem through the Linked Data (LD) paradigm. In this approach, all information is encoded as triples consisting of a predicate and two arguments (usually called the `subject' and the `object'). This resolves the syntactic interoperability problem, since all databases then have the same structure: a big table of triples. The triples can be encoded in different ways, but RDF is the most used encoding mechanism, and it is also used in WP4. Semantic interoperability is addressed by harmonizing the vocabularies used and ensuring that the elements from the triple (the predicate, the subject and the object) are associated with clearly defined concepts. By turning to the LD paradigm, links can also be made with external data sources encoded as linked data, and that is already a huge collection and it is continuously growing.

By encoding all data as triples relations can be sought across different databases, possibly from different levels, which can be used to test hypotheses about correlations that could not be investigated before, and by data mining the combined LD databases new correlations may be found.

Searching and analysing the data in LD require a special query language. Such a language exists (SPARQL), and CLARIAH will experiment with this query language and its suitability for making queries in the social economic domain.

Since all information is encoded in triples, which have a very small granularity, one needs a huge number of triples to encode all information. This, in its turn, imposes special requirements on storage and on systems that enable efficient search in such large sets of triples. Research in these matters is also carried out in CLARIAH. The major data centre for WP4 is the International Institute for Social History.

WP5

In the media studies work package (WP5) the most dominant data type is audio-visual data, such as films, TV-programmes, radio broadcasts, vlogs, recorded interviews, etc. WP5 develops a software suite that enables search in and analysis of such data, in combination with information sources of different types such as natural language texts in newspapers, journals and the new social media (Twitter, Facebook, blogs, etc.). Audio-visual data are individually large (orders of magnitude larger than textual and structured databases) and require special tools for viewing, browsing and searching. Moreover, because of copyright restrictions, access to audio-visual data requires specific authentication measures. The media suite will provide access to relevant audio-visual collections, collections providing context to audio-visual material, and tools. It is built by integrating (partially redesigned) components of a set of existing tools (DIVE, AVResearcherXL, TROVe, Oral History Today) developed in earlier projects. The major centre for WP5 is the Netherlands Institute for Sound and Vision.

WP2

The discipline-independent WP2 develops software and data that are needed as part of the generic infrastructure. It creates new and connects existing databases with information about persons, locations, and documents, and it creates a database in which words (terms) are linked to concepts as a function of time, so that words occurring in texts can be interpreted in the sense(s) they had at the moment they were written down. It will also make it possible to track meaning changes over time, and will enable concept search through time not hindered by different terms used in different times.

WP2 also uses the LD paradigm, and actively investigates linking its own resources to external knowledge resources such as dbPedia. The major data centres for WP2 are Huygens Institute and the Institute for the Dutch Language.

Cross-WP data

In the CLARIAH-CORE project we will experiment with techniques to extract information from one data type and convert it into a different data type. For example, we will carry experiments to extract information hidden in natural language texts and turn them into structured data. One experiment is running in the Athena subproject on extracting information on fauna and flora from textual documents (WP3 / WP4 cooperation). Another is planned for extracting structured data from pictorial and textual data on 'filmladders' (WP3 / WP5 cooperation).

Metadata

A special type of data is `metadata', technically meaning *data about data*. Though the distinction between data and metadata may look simple, in practice it is not always so easy to distinguish them. On the other hand, it is also often not so important whether data are treated as data or as metadata, as long as it is dealt with. CLARIN uses a specific framework for metadata, called CMDI (Component-based MetaData Infrastructure), around which many tools and services have been built. These include a metadata profile and component registry and editor, various metadata editors, and a large number of predefined components and profiles. The Virtual Language Observatory harvests all the metadata provided by CLARIN centres and makes them (and through them the data they describe) findable and accessible. Since the LD framework is gaining in weight internationally, and is also heavily used in CLARIAH-CORE, a special project is underway to convert CMDI-metadata into LD and vice-versa.