Document: http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-en.pdf, modified: 2009-03-27

This document explains the LETEC (and TEACHING Corpus, extension of olac-linguistic-type) extension to OLAC (*Open Language Archives Community*, http://www.language-archives.org)

Notion of LETEC; Learning and TEACHING Corpus

Definition

We define a *Learning & Teaching Corpus* as a structured entity containing all the elements resulting from a communicative on-line learning situation, whose context is described by an educational scenario and a research protocol. The core data collection includes all the interaction data, the productions of the course participants, and the tracks, resulting from the participants' actions in the learning environment and stored according to the research protocol. In order to be able to be shared, and to respect participant privacy, these data should be anonymised and a license for its use be provided in the corpus. A derived analysis can be linked to a given set of data under consideration, used or computerized for this analysis. An analysis consisting in data annotation/transcription/transformation, accurately connected to its original data, can be merged with the corpus itself, in order for other researchers to compare their own results on a concurrent analysis or to build their complementary analysis upon these results.

The definition of a Learning & Teaching Corpus as a whole entity comes from the need of explicit links, between interaction data, context and analyses. This explicit context is crucial for an external researcher to interpret the data and to perform its own analyses.

This definition seeks to capture the context of the data stemming from the course in order to allow a researcher to look for, understand and connect this information whether or not he/she was involved in the original course.

Corpus composition and structure

The main components of a learning corpus (see figure 1) are:

- The Instantiation component, the heart of the corpus, which includes all the interaction data, the production of the online course participants, completed by some system logs as well as information characterizing the participants' profile.
- The Context concerns the educational scenario and the research protocol (optional element).
- The License component specifies both corpus publisher's (editor) and users' rights and the ethical elements toward the course participants. A part of the license component is private, held only by the person in charge of the corpus. Only this private part may contain some personal information regarding the course participants.

- The Analysis component contains global or partial analysis of the corpus as well as possible transcriptions.

The Mulce structure aims at organizing the components of the corpus in a way that enables linking the components together. For example a researcher, while reading a chat session (which belongs to the instantiation component), must have access to the objectives of the activity (which belongs to the pedagogical context).

Moreover, it is important for the Mulce format to allow data-mining of the component data on the platform (cf. for these two points the consultation section of the corpora). A standard exchange format is also required to download the whole corpus. Considering these constraints, we chose the IMS-CP formalism (IMS-CP 2004) as the global structure. This XML formalism fits these constraints by expressing metadata, different levels of description, and an index pointing to the set of heterogeneous resources. Each corpus is thus archived as a Content Package (IMS-CP 2004), including metadata, descriptions and related resources used in each of the components.

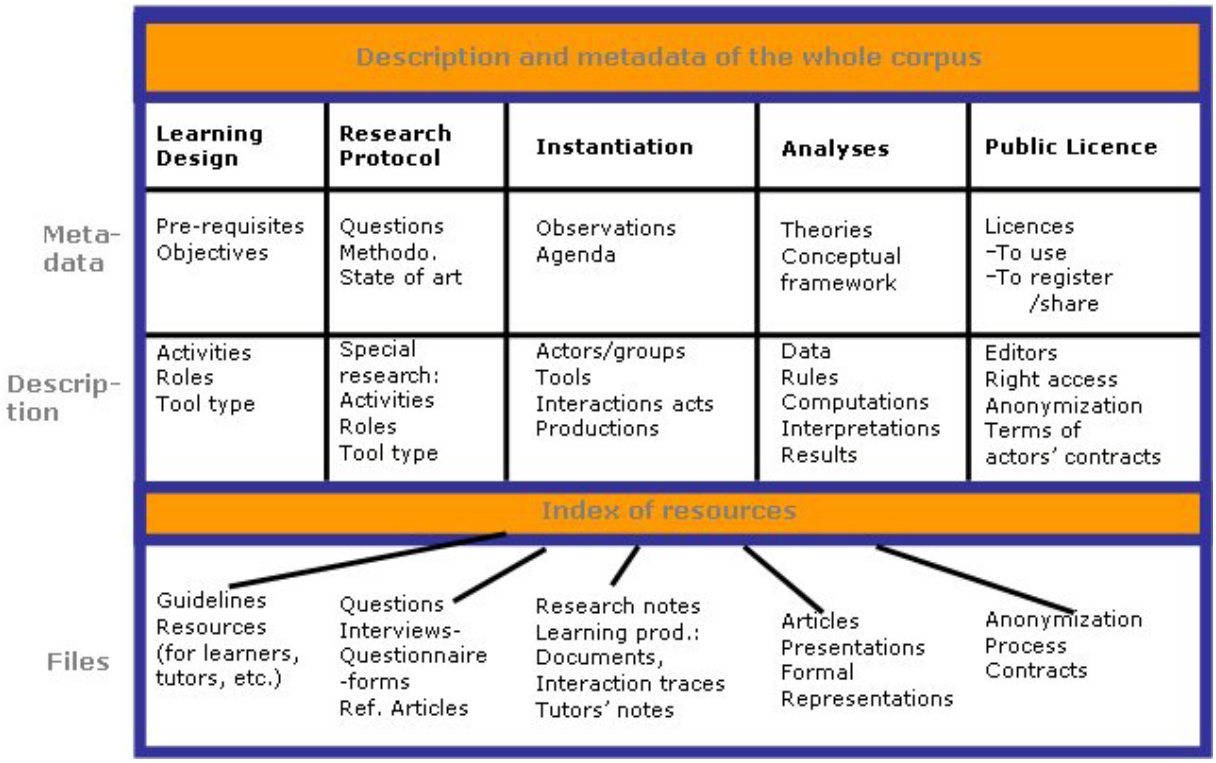

Figure 1. Teaching and learning corpus: the main components in a Content Package.

Instantiation formalism: Participants and environment description

The instantiation phase consists in describing (1) the participants, (2) the technological environments, (3) the tools used during the learning activity and (4) the groups and their members. We consider that the pedagogical scenario can describe the generic activity of a group by specifying the roles without assigning them to participants and declaring only the type of the involved tools. For example, in the abstract pedagogical scenario, one can define a negotiation activity for the production of a collective document that has to be performed by each group using a chat and a forum. In the instantiation part of the corpus, we have to define the participants involved and concrete environment used. For example, the main environment used may be a Learning

Management System, whose chat and forum tools have specific features (speech acts, attached file, etc…). This description results in the definition of the environment feature together with the specification of the structure of the traces collected during the learning activity. Participants' general description concerns their age, gender, institution and some of cultural attributes if needed (country, mother tongue, etc.) When more specific information is required, the structure may be extended by a specific XML schema.

Instantiation formalism: Workspace concept

The hierarchical structure of the learning stage (potentially spread in parallel groups) is captured in the Workspaces element, i.e.: a sequence of "workspace" elements (see figure 3). A workspace is generally linked to a learning activity (of the pedagogical scenario). It encompasses all the events observed during this activity, in the tool spaces provided for this activity, for a given (instantiated) group of participants. As shown on figure 2, a workspace description includes its members (references to the participants registered for the learning activity), starting and ending dates, the provided tools and the traces of interaction that occurred in these tools. In order to fit the hierarchical structure of learning and support activities, a workspace can recursively contain one or more workspace elements.

The lists of places, sessions, descriptors, contributors and sources defined in the workspaces element can be referenced by workspace, contribution, or act elements. For example, descriptors may list identified categories so that each act of the acts element list could refer to one or more of these categories. This principle enables to browse the interaction data in many different ways, independent to the concrete storage organization in the XML document.

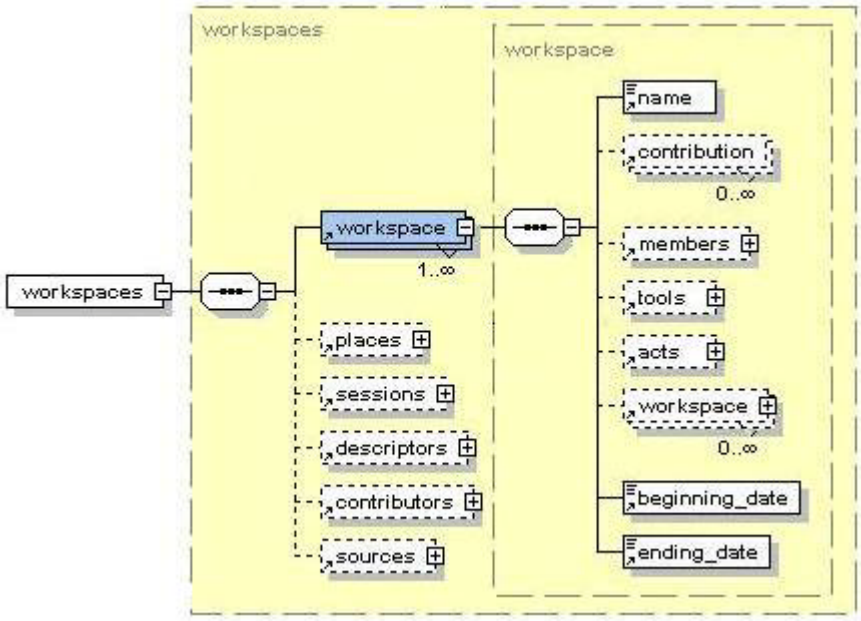

Figure 2. Extract of the XML Schema.

Our specification describes communication tools and their features with a great level of precision. The corpus builder can specialize/particularize the schema (i.e., restrict it) to fit the specific tools and features proposed to the learners in a specific learning environment. In the meantime, if a tool cannot be described with the specification, one can augment the schema by adding new elements, in

order to take into account the tool's specificities. Both of these mechanisms offer two ways, the specification can be extended to fit the analysis needs.

Moreover, recursive workspace description enables the corpus descriptor to choose the grain at which he needs to describe the environment. Thus, a workspace can be used to describe a complete curriculum, a semester, a module, a single activity or a work session (a concept generally related to synchronous learning activities). The workspace concept represents the space and time location where we can find interaction with specific tools. This notion has the same modularity as the EML learning units (EML, 2000).

Devices and tools within which interaction occurs can be as different as a forum, a chat or collaborative production tools (e.g., a conceptual map editor, a collaborative word processor, a collaborative drawing tool).

Interaction traces are stored according to the act's structure presented on figure 4. All actions, wherever they come from are described by an act element. An act necessarily refers to its author identifier (defined in the members list – figure 2), and a beginning date. Depending on the nature of the act (act_type), an optional endind_date can be described. The act_type element is a selector. The actual content (or value) of the act depending on its type, is stored in the appropriate structure

References

Reffay, C, Chanier, T., Noras, M. & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. In Basque, J. & Reffay, C. (dir.), *numéro spécial EPAL (échanger pour apprendre en ligne), Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF),* 15, [http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01p.pdf, http://edutice.archives-ouvertes.fr/edutice-00159733]

Mulce (2009). French version of this document [http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-fr.pdf]

Reference and use of this document :
Mulce (2009). Notion of LEarning and TEaching Corpus. [http://mulce.univ-fcomte.fr/metadata/mce_LETECorpus-en.pdf]

Mulce project (ANR-06-CORP-006) supported by ANR AGENCE NATIONALE DE LA RECHERCHE