Sharing corpora and tools to improve interaction analysis

Christophe Reffay[1], Marie-Laure Betbeder[1]

[1] LIFC: Computer Science laboratory of the University of Franche-Comté
16 Route de Gray
25030 Besançon cedex, France
{Christophe.Reffay, Marie-Laure.Betbeder}@univ-fcomte.fr

Abstract. A very wide range of online interaction analysis staying in the hands of researchers, and tools being implemented in research prototypes, often used only in non-replicated experimentations, we point out the need for TEL research community to reach large scale validation for its results. This paper is a concrete step in this direction. For a deeper collaboration in our community, we suggest to share structured data collections. The Mulce project aims at proposing a structure for teaching and learning corpora (including pedagogical and research context), and especially for interaction tracks. Two main corpora are built according this structure. This paper defines a teaching and learning corpus, shows its main structure and browses some parts of the structured interaction data. We also describe the platform that enables the community to browse and analyze a shared corpus.

Keywords: e-research, corpora sharing, interaction analysis.

1 Motivation

In the last twenty years, we saw the emergence of an incredible number of tools, services and platforms. One technology quickly replaced the previous one, offering more and more potential for interaction analysis. Some voices in our communities are pointing out the problem of little impact of our research outcomes on real learning situations: our very intelligent tools and services often stay in the researchers' hands and rarely go beyond the prototype stage. The time rate for technology innovation is too high, comparing to the time needed by social science to validate some of our prototypes.

In this paper, in order to propose to the community a way to access, share, analyze and visualize learning and teaching corpora, we propose a new formalism [1] which defines, describes and structures data provided by on-line training. Before presenting our proposal, in this section, we come back to the validation of indicators and tools for Technology Enhanced Learning and present other works related to this contribution.

The study of collaborative online learning, whether aimed at understanding this form of situated human learning, at evaluating relevant pedagogical scenarios and

settings or at improving technological environments, requires the availability of interaction data from all actors involved in such learning situations, including learners and teachers.

We can find a lot of technical proposals for indicators for social or cognitive process monitoring especially in the TEL, Intelligent Tutoring System and Computer Supported Collaborative Learning communities of the last decade. If some of these indicators are very specialized, i.e. strongly related to a given tool or activity, we find also very general purpose indicators taking their raw data from widely used communication tools like text chat, text conferencing or e-mail.

These technical implementations for indicators conceptually provide a large range of possibilities and make this research area very creative. The very most part of these indicators (including ours) are designed in a given context, where they show some interesting properties and even promise usefulness for the various actors involved in real situations. Unfortunately, these indicators often stay in the researchers' hands and are rarely used by real actors of the situation. As far as we know, none of them have been validated or at least evaluated by real/concrete actors. The need of validation for these indicators, at least in a given context, becomes crucial if we want this domain to contribute to the real world distance learning area.

These indicators are also rarely reused in other situations or contexts. We argue in this paper that our research community should be able to widen the validity of an indicator by testing it on different situations.

In their work [2] on coding and counting analysis methodology, the authors already pointed out the weakness of our research domain. Replicability, reliability and objectivity need to be improved in our work.

The main idea of research collaboration is already well expressed in [3] in the following terms:

"There is urgent need of putting together complementary strengths and contexts and combining our insights as rapidly as possible to make a greater impact and further elevate our research quality at the same time. Research generally has had a small voice in national educational outcomes; we can speak louder if we speak together."

We know how hard it is to build natural classroom situations, called here authentic learning situations. This is one of the reasons why we launched the Mulce [4] project. Instead of having hundreds of unclassified learning situations, where the data of each are available only by the researchers that built it, we argue that our communities would gain maturity and deepen its understandings by sharing some of the representative situations. Such data could be used as a test-bed for the variety of indicators or methods to analyze various facets of the collaboration.

For the Intelligent Tutoring System (ITS) field, the PSLC DataShop [5] presented in [6] provides a data repository including data sets and a set of associated visualization and analysis tools. These data can be uploaded as well-formed XML documents that conform to the Tutor_message schema. The goal is to improve ITS the data are logged from. The data sets are fine-grain mainly automatically generated by the ITS themselves and focus on action/feedback interaction between learners and (virtual) tutor tools.

In the CSCL community, a very interesting framework: DELFOS [7, 8] provides similar proposals as the Mulce project. It defined an XML based data structure [9] for collaborative actions in order to promote interoperability (between analysis tools), readability (either for human analysts and automated tools) and adaptability to different analyzing perspectives. Some of these authors joined the European research project (JEIRP–IA) on Interaction Analysis and reported in [10] a template describing IA tools and a common format. This common format should be automatically obtained from Learning Support Environments (by an XSL transformation) and either directly processed by new versions of Interaction Analysis tools, or automatically transformed in their original data source format to be processed by previous versions of theses tools. The resulting common format focused more on technical interoperability than on learning context or human readability. The context is given for fine grain interaction.

A very interesting experience in the CSCL community has been initiated by the Virtual Math Team [11]. Multimodal Chat sessions (namely teams B and C of the 2006 Spring Fest) in the Virtual Math Forum have been collected and delivered to numerous (28) external collaborators coming from 11 countries, 18 institutions and 8 different research fields. Every collaborator applied his/her own analysis methods and tools processing these interaction data in order to see what came up. The result is reported in [12]. The same data set has been used also for a pre-Workshop of the last CSCL conference in Rhodes. In this context, we showed how this data set can be structured in a Mulce structure and a new collaboration is currently building this data set as a fully documented corpus to be available in the Mulce repository.

In the Mulce structure, the learning situation and the research context are described as wholes possibly in different formats (IMS-LD, LDL, MotPlus, simple text document, etc.) If they conform to IMS-LD, their identified included objects can be referred to by the workspace elements structuring acts' lists that are recorded in the instantiation part. The nature of sharing perspectives is very different: in the JEIRP, the goal is to share a schema structure, whereas the Mulce platform's main objective is to share the data collections.

For this last issue, an impressive work has been done in the Dataverse Network project [13] described in [14]. We agree with the members of this project on the fact that datasets have to be made available, or at least identified and recorded in a fixed state in order to make sure that data used for a given publication are the same as those identified and (hopefully) made available for other researchers. We also consider that such a (data) publication, when connected to a traditional paper published in a journal or conference, would increase the value of this article and of the related journal (or conference proceedings).

In the Mulce project, we provide a technical framework to describe an authentic situation, described by a formal or informal learning design or detailed guidelines, with a representative number of actual participants, according to a research protocol. We also: define a "Learning and teaching Corpus", provide a technical XML format for such a corpus to be sharable and we are currently developing a technical platform for researchers to save, browse, search, extract and analyze online interactions in their context. The main idea of the Mulce project is to provide contextualized interaction data connected to published results.

Considering today's available technology, Lina Markauskaite and Peter Reimann drew an ideal research world in [15] where grid computing, middleware services, tools managing remote resources, open access to publications and data repositories, open and interactive forms of peer review process, constitute great potential for e-research. We globally share the same vision for the future of research. Even if we consider that the way to reach this ideal vision is rather long, the main contribution of this paper can be considered as a modest but concrete step in this direction by giving a definition and the data structure of a teaching and learning corpus as well as the associate platform to share such corpora.

Availability of data should enable deeper scientific discussion on previously published results. Other researchers may be able to verify or replicate the methods proposed. It becomes possible to compare methods on the same data and then discuss the result or the efficiency of the methods. This way, different analyses can be done on the same set of interaction data. The Mulce platform currently plans to connect these analyses to the set of data they are based on. Such a set can be part of one or more corpora available on the Mulce platform.

Even if sharing research data on collaborative learning has a wider range of implications, in this paper, the discussion, introduced in the next section, focuses on the validity of indicators given in the TEL literature. The main contribution, presented later on, is the open Mulce XML structure for interaction data and the related Mulce platform allowing our communities to share such data collections. We will first define a teaching and learning corpus, then describe its components, and detail the structure of interaction data. The Mulce platform design is then presented in section entitled proposal, before a brief conclusion.

2 How to validate our tools or indicators?

In the review [16] of the "*state of the art technology for supporting collaborative learning*", we find 23 referenced systems allocated to 3 main categories: Mirroring tools, Metacognitive tools and Guiding systems. These are allocated to the categories according to the locus of processing (i.e. where diagnosis and remediation are synthesized). As mentioned by the authors in their conclusion, "*We have not yet seen full-scale evaluations of the types of systems we have covered here*" and "*If our objective is to assist students and teachers during real, curriculum-based learning activities, we must also understand how well our laboratory findings apply to natural classroom situations.*"

In this very nice classification, and particularly for both of the last categories, the definition of a "desired" or "ideal" state of interaction is crucial. When should the supervisor (in the second category) or the system (in the third one) consider that the current state is too far from the desired state? This decision can be made by a simple comparison with a threshold like the desired number of messages suited during a certain period of time, or the number of group members one learner has interacted with…

Sometimes, like in [17], this threshold can be directly extracted from the learning design, where the guidelines for learners explicitly indicate a list of precise tasks involving a countable number of interaction messages.

For sessions (replicated over years) applying the same learning design to similar cohorts of learners, we can use a first session to calibrate the "ideal state". For example, in an English as second language acquisition module, the online Copeas experiment has been used to measure the time each of the actors (tutor and learners) has talked during the online audio-graphic synchronous sessions [18]. These measures are related to the different profiles of learners: English level, age, favorite modality of interaction [19], etc.

For metacognitive or guiding tools currently designed and implemented for further learning sessions, a set of representative interaction data collections would be very useful (if available) for a calibration step. In such a case, these tools could be tested in the design process by using available (shared) data collections and be applied and evaluated directly by real actors during the first experiment.

We can quote [20] as a good example of experiment where mirroring tools are actually tested by learners to get a bird-eye on their ongoing collaboration in a long-term project using a wiki. In their paper, the authors conclude that this first step of tool evaluation showed its usefulness especially for group leaders, and had a positive effect on collaboration management. A better understanding of the representation seems to be needed by learners to improve their interpretation. The authors plan to give more control to users to choose what and how information should be visualized in order to get a better appropriation of the tool. As wiki has become popular in our research experiments, we could imagine that other researchers have similar tools or analysis methods that run on such data. Their availability could help to compare these tools if they have the same goal, or to enrich the analysis if they give a complementary point of view.

For computer scientists, it could be enough to put the raw data of the wiki logs and contents on a shared repository, but for the major part of our communities that would analyze the content and draw some interpretations of these analyses, the format of the data should be understandable and the context of the situation readable.

Either we can keep developing more and more prototypes giving intelligent feedback to their (hypothetic) users. This way implies that a great part of our force is dedicated to the construction of new experiments for most of our new prototypes. We can try to reuse a very interesting analysis tool in a slightly different context, but, in the worse case, with very different data formats.

Or we could try to share some representative authentic learning sessions, for a wider use in a test-bed platform, involving researchers from a wider range of sciences, sharing their complementary analysis. Even if some innovative experiments will remind necessary, a lot of time for a lot of us could be dedicated to deepen understanding, to compare and to validate thresholds values, analysis methods and tools, and to build large scale validation of them.

In other words, the questions behind are: What is more efficient between sharing data and sharing format or tools (without data)? Whom for?

The rest of this article is a proposal of "how to share" such a data collection. The next section defines the learning and teaching corpus and describes the structure of its main components.

3 Proposal

Our proposal consists in (1) a formalism to describe learning and teaching corpora and (2) a platform to share these corpora [1]. The formalism defines the information which can be contained in a corpus and the structure of the data. Through the platform, researchers can share their corpora with the community and access the data shared by other members of the community.

To share a corpus, a researcher has to provide metadata describing the corpus' components and upload a file describing each component. While accessing a corpus, an identified researcher is provided with a variety of tools to browse the corpus components, to navigate through the contextualized interaction data, to visualize and to analyze them.

3.1 Proposal 1: Learning and teaching corpus formalism

In the many fields involved in computer mediated interaction analyses, we can find different research methodologies that result in different needs and especially different ways to save and describe the data. If the definition of "learning and teaching corpus" necessarily depends on the way research experimentations are conducted, we claim that our definition is general enough to fit this variety and a crucial point for the concrete structure is to make explicit the methodological choices for a given experiment.

In this section, we first present the main phases involved in this methodological process. Then, we give the derived definition of a "learning and teaching corpus" and explore the structure of its main components.

Building and recording interaction in an online training
A general organization for an online experiment is illustrated on figure 1.

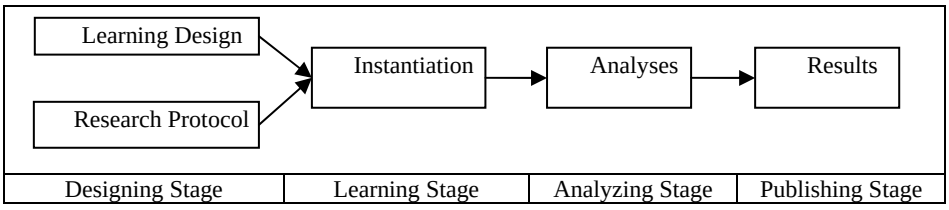

Fig. 1. Building a research experiment for an online training: chronology.

In a first stage, the educational scenario is described, at abstract level, by defining the educational prerequisites and objectives, the abstract roles (learner, tutor, etc.), the learning activities and the support activities with their respective environments (abstract tools, e.g. chat, forum, etc.) When the training has to be observed for a research study, the researchers define on the one hand the research questions and objectives and on the other hand the list of observable events to be logged. This

documentation makes explicit the research protocol or context of the experiment: i.e.: what will be evaluated, are there pre- or post- tests, or training interviews? In the second stage, the training actually takes place. The abstract roles (designed in both parts of the first stage) are endorsed by real actors, and abstract environments have been implemented in particular platforms including identified tools. This is the instantiation phase where embodied learners and tutors actually run the activities and identified researchers collect their observable actions (interactions and productions). Specific activities designed in the research protocol may also take place during this period: e.g. pre- or post- tests, interviews, etc. At the end of the training, i.e. when learners and tutors are gone, the collected data can be structured and analyzed by researchers. These analyses hopefully lead to research publications that summarize the context and the methodology and try to explain the results. The data collection is not disseminated.

Both documentations of the design phase describe the context of the experimentation. This information often stays in the head of the researchers involved in the experimentation. Instantiation phase produces the core data collection that is analyzed in the third stage. Having the context in mind, these researchers can interpret properly their results during the analysis phase.

As a consequence, in order to make this data collection sharable with external researchers, we show how the various phases presented above become the main components of the corpus defined in the next section.

Learning & Teaching Corpus: Definition

We define a Learning & Teaching Corpus as a structured entity containing all the elements resulting from an 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 training actors' production, and the tracks, resulting from the actors' actions in the learning environment and stored according to the research protocol. In order to be sharable, and to respect actor privacy, these data should be anonymised and a license for its use be provided in the corpus. A derived analysis can be linked to the set of data actually considered, used or computerized for this analysis. An analysis consisting in a data annotation/transcription/transformation, properly connected to its original data, can be merged in the corpus itself, in order for other researchers to compare their own results with a concurrent analysis or to build their complementary analysis upon these previous shared 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.

The general idea of this definition intends to grasp the context of the data stemming from the training to allow a researcher to look for, understand and connect this information even though he has not attended the training course.

Corpus composition and structure

The main components of a learning & teaching corpus (see Fig. 2) are:

- The Instantiation component, the heart of the corpus, which includes all the interaction data, production of the on-line training actors, completed by some system logs as well as information characterizing actors' profile.
- The Context concerns the educational scenario and the optional research protocol.
- The License component specifies both corpus publisher's (editor) and users' rights and the ethical elements toward the actors of the training. 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 actors of the training.
- The Analysis component contains global or partial analysis of the corpus as well as possible transcriptions.

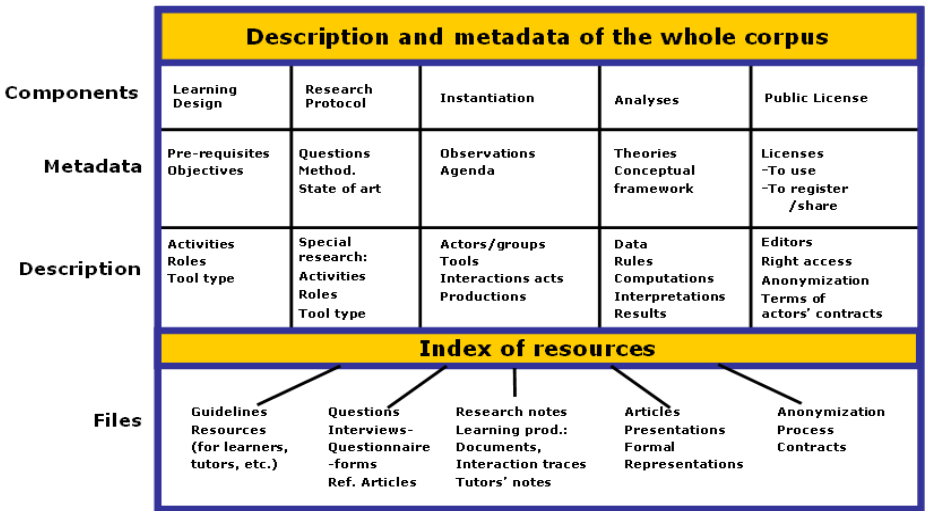

Fig. 2. Teaching and learning corpus: the main components in a Content Package.

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 be able to read the objectives of the activity (which belongs to the pedagogical context).

Moreover, it is important for the Mulce format to allow, digging the component data on the platform (cf. for these two points the section entitled "Browsing and analyzing corpora"). A standard exchange format is also required to download the whole corpus.

Considering these constraints, we chose the IMS-CP formalism [21] as the global container. 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 [21], including metadata, descriptions and related resources used in each of the components.

Instantiation component: Actors and environment description
This component consists in describing (1) the actors, (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 actors 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 all actors involved and concrete environment used. For example, in the activity described previously, the main environment used was the WebCT platform, 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 tracks collected during the learning activity. Actors' general description concerns their age, gender, institution and some of cultural or cognitive profile attributes if needed (country, mother tongue, etc.) When more specific information is required, the structure may be extended by a specific XML schema.

Instantiation component: Workspace concept
The hierarchical structure of the learning stage is captured in the *workspaces* element, i.e.: a sequence of *workspace* elements (see Fig. 3).

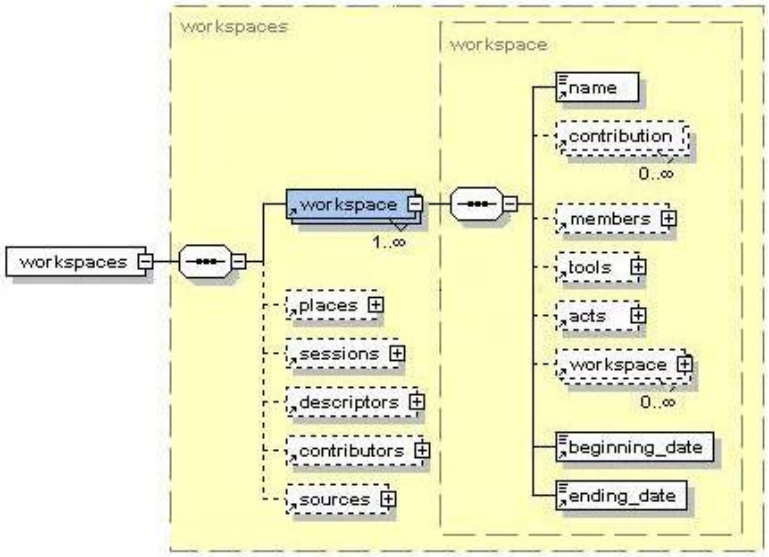

Fig. 3. Extract of the XML Schema.

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 actors. As shown on figure 3, a workspace description includes its *members* (references to the actors registered in the learning activity), starting and ending dates, the provided *tools* and

the tracks 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.

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 tools specificities or 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 identified tools. This notion has the same modularity as the EML learning units [22], [23].

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).

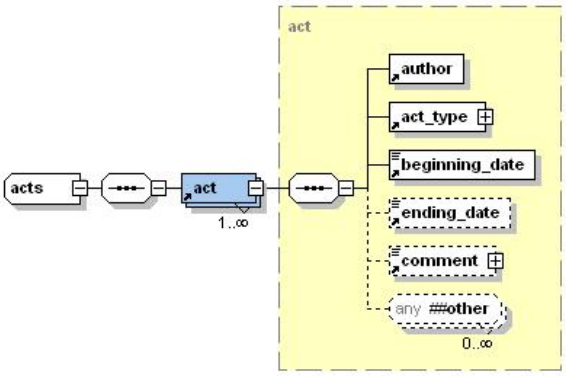

Fig. 4. Extract from the XML Schema – the act concept.

Interaction tracks 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 – Fig. 3), and a *beginning_date*. Depending on the nature of the act (*act_type*), an optional

endind_date can be specified. 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.

For example, a chat act (see Fig. 5) can have the type in/out (actor entering/leaving), it may contain a *message*, can be addressed to all the workspace members or to a specific one (e.g. if it is a private message). A chat act can contain an attached document (*file*) which in turn is described by a name, a type and a date.

Optional element comment contains a sequence typed text of any type and can be used to store researchers' annotations. The last optional element of the *act*'s structure (any) leads to any extension not provided in our schema.

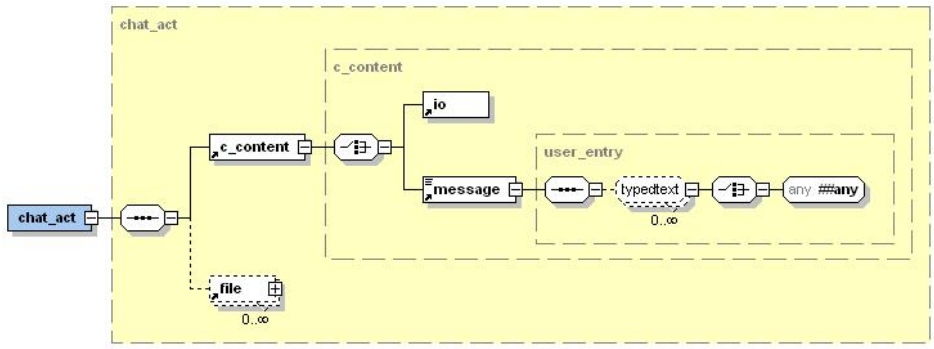

Fig. 5. Extract from the XML Schema – the chat act concept.

This XML Schema defines the storage structure for many act types, e.g.: forum message, chat act, transcribed voice act, and more. For lack of space, this paper only gives some of the main ideas of this schema, but the complete schema for structured information data is available online [24].

The definition, composition and structure of a Learning & Teaching Corpus have been presented in the sections above. The next one explains how these data structures can be shared and computerized on the Mulce platform.

3.2 Proposal 2: a Platform for corpus sharing

Sharing corpora

Once data have been collected, structured and described by metadata, we are ready to share them on the Mulce platform. Being connected with other Open Archive Initiative repositories [25] [26], the Mulce platform deals with sharing metadata and our corpus objects become visible for the whole community. Two mains corpora (Simuligne and Copeas) are already uploaded. About twenty related corpora containing analyses are also in our repository. This paper is also an invitation for all researchers to prepare their corpora in order to share them on the Mulce platform, keeping them readable.

The deposit of a corpus consists in declaring it, describing it by means of general metadata, and uploading its components (described previously). Each component has a specific formalism. These can either be standard formalism such as Learning design [27] (used for the context components: educational scenario and research protocol), or the specific formalism described here above for structured interaction data. If these recommended formalisms are used to describe the various components of the uploaded corpus, the researchers will fully benefit from all the tools provided one the Mulce platform to navigate and analyse the entire corpus. Otherwise, the corpus will be downloadable as is by other researchers. Each component is described by its specific metadata. On the Mulce platform, these metadata can be used by a researcher to find corpora that fit particular constraints. For example the researcher can select the corpora pertaining to its own research interests, either in terms of used tools, of targeted audience or logged tracks.

Browsing and analyzing corpora

The second part of the platform proposes the visualization, the navigation and the analysis of structured interaction data. Corpora or selected parts of corpora can be downloaded by identified researchers. In this part, two distinct aspects are considered: the navigation / visualization aspect, and the analysis aspect of corpora.

The interest of the navigation / visualization aspect is twofold. Firstly, the corpus becomes independent to the (evolving) software, where originally interaction took place. This is a major benefit for data longevity and reusability. Secondly, because of the main attention paid on the context of interaction in the Mulce project, the interaction navigator makes explicit links between interactions and their surrounded context. Finally, the researcher can select a part of a corpus by means of requests. He can, for example, select all the interactions of an actor using a specific communication tool. For each of the interactions he can access to the prescribed educational activity.

We are currently developing a user interface enabling navigation through different corpora. A first form provides a selection of corpora according to the following criteria: participants (students, tutors, native speakers), technologies (asynchronous LMS, audio-graphic conference, discussion forum, chat, …), pedagogical dimensions (global simulation, intercultural scenario, English and ICT, …), learning fields (French as foreign language, English for ICT, …), analysis tools (forum analysis, synchronized multimodal layouts, social networks analysis, …) language used, interactions and modalities (spatial-, spoken-, textual-, iconic-, multimodal scaffolding language, ..) The result of this request is a list of corpora matching the criteria, with synthetic information. Once selected, a corpora can be described (metadata), browsed (each component with its specific interface), or scanned in order to select or highlight particular acts.

The analysis aspect of corpora concerns the use of tools based on the instantiation component formalism. As an example, patterns of interactions can be detected by a pattern discovery tool [28]. The XML format being defined, we hope that different analysis tools (including indicator synthesis), coming from various teams, will have a version that can operate on the Mulce structure. Tools can be either integrated to the platform for an online use or downloaded from the platform for an offline use. The tools proposed on the platform will originate from our research team or from partnership. For example we have two running collaborations: the Calico project and

Tatiana. The Calico project ([29], [30]) aims at proposing different visualization and analysis tools [31], specialized on discussion forums. Tatiana ([32], [33], [34]) includes a navigator, a replayer and an annotator. The replayer functionality synchronizes the various data sources and ... “aims at bridging the gap between having the data of an experiment and being in the flesh of the observer” [32]. We are currently adapting its XML schema to fit ours and extend its visualization functionalities to other communication tools. We are interested in other collaborations aiming at providing other analysis tools.

Technology

The Mulce platform is developed over a Java/JEE application stack running on the Tomcat servlet server [35]. The implementation conforms to the MVC2 Design Pattern, using the Struts framework [36]. This application provides a single point of control and then facilitates security concerns. In order to get independence between our core computing process and the database, we consider a mapping by using the Hibernate framework [37]. Finally, the Graphical User Interface takes advantage of the SiteMesh frameset [38].

Because our field (linguistics) already owns its Open Archive Implementation (OLAC), we chose to connect our server (as a data repository) to the OLAC harvesting network (compliant to the OAI-PMH protocol).

4 Conclusion

Joining the voice of other researchers, this paper deals with the problem of TEL research impact and focuses on the methodology to validate indicators and analysis tools provided by our communities. Because research (not only learning) also could benefit from collaboration tools on the Internet, we think that a more collaborative research could have a greater impact on indicators and then, on real world online learning. Due to the fact that experiments in online learning involve human beings, embarking their specificities and cultural context, the replication is very hard to achieve. This problem prevents two essential validation processes. Two concurrent indicators, used in two different contexts, cannot be compared. And, because original interaction data are not available for other researchers, none of the indicators can be tested on external experiments. This leads to a lack of large scale evaluation for each indicator or tool.

In order to concretize a first step towards e-research, the Mulce project aims at sharing contextualized interaction data in a Learning & Teaching Corpus. Sharing corpora means building a test-bed to compare our indicators and analysis tools on fixed data. This paper proposes a definition, a composition and a structure for such a corpus. A related platform is currently implemented to share, browse and analyze shared corpora.

References

1. Reffay, C., Chanier, T., Noras, M., Betbeder, M.-L.: Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. In: STICEF journal (Sciences et Technologies de l´Information et de la Communication pour l´Éducation et la Formation), vol. 15, 25 p. (2008)
2. Rourke, L., Anderson, T., Garrisson, D.R., Archer W.: Methodological Issues in the Content Analysis of Computer Conference Transcripts. In: IJAIEd, vol. 12, pp. 8--22 (2001)
3. Chan, T., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., Patton, C., Cherniavsky, J., Pea, R., Norris, C., Soloway, E., Balacheff, N., Scardamalia, M., Dillenbourg, P., Looi, C., Milrad, M., Hoppe, U.: One-to-one technology-enhanced learning: An opportunity for global research collaboration. In: Research and Practice in Technology Enhanced Learning, vol. 1(1), pp. 3--29 (2006)
4. Mulce: French national research project 2006-2010 (ANR-06-CORP-006), coordinated by T. Chanier. http://mulce.univ-fcomte.fr/axescient.htm#eng
5. The Pittsburgh Science of Learning Center (PSLC) DataShop: https://pslcdatashop.web.cmu.edu/
6. Koedinger, K.R., Cunningham, K., Skogsholm, A.: An open repository and analysis tools for fine-grained, longitudinal learner data. In: Proceedings of the First International Conference on Educational Data Mining, pp. 157--166 (2008) http://www.educationaldatamining.org/EDM2008/uploads/proc/16_Koedinger_45.pdf
7. Osuna, C.: DELFOS: A Telematic and Educational Framework based on Layer oriented to Learning Situations. PhD Dissertation, Universidad de Valladolid, Valladolid, Spain (2000)
8. Osuna, C., Dimitriadis, Y., Martínez, A.: Using a Theoretical Framework for the Evaluation of Sequentiability, Reusability and Complexity of Development in CSCL Applications. In: Proceedings of the European Computer Supported Collaborative Learning Conference, Maastricht, NL, march (2001)
9. Martinez, A., De la Fuente, P., Dimitriadis, Y.: Towards an xml-based representation of collaborative action. In: proceeding of Computer Supported Collaborative Learning conference (CSCL), Bergen (2003) http://hal.archives-ouvertes.fr/hal-00190429/fr/
10. Martinez, A., Harrer, A., Barros, B.: Library of Interaction Analysis Tools. Deliverable D.31.2 of the JEIRP IA (Jointly Executed Integrated Research Project on Interaction Analysis Supporting Teachers & Students' Self-regulation). KaleidoScope (2005) http://www.rhodes.aegean.gr/ltee/kaleidoscope-ia/Publications/D31-02-01-F%20Lirbary%20of%20IA%20tools%20.pdf
11. Virtual Math Team: http://www.ischool.drexel.edu/faculty/gerry/vmt/index.html
12. Stahl, G.: Studying Virtual Math Teams. New York, NY: Springer (2009)
13 The Dataverse Network Project: http://thedata.org/
14. King, G.: An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. In: Socialogical Methods & Research, vol. 36(2), pp. 173--199 (2007) http://gking.harvard.edu/files/dvn.pdf
15. Markauskaite, L., Reimann, P.: Enhancing and Scaling-up Design-based Research: The potential of E-Research. In: Int. Conference for the Learning Sciences, ICLS'2008. Utrecht, NL, June, 8 p. (2008)
16. Soller, A., Martinez, A., Jermann, P., Muehlenbrock, M.: From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. In: IJAIED vol. 15, pp. 261--290 (2005)
17. Reffay, C., Chanier, T.: How social network analysis can help to measure cohesion in collaborative distance-learning? In: proceeding of Computer Supported Collaborative Learning conference (CSCL'2003), Bergen, pp. 343-352 (2003)

18. Vetter, A., Chanier, T.: Supporting oral production for professional purpose, in synchronous communication with heterogeneous learners. In: The journal of Computer Assisted Language Learning. Recall, vol. 18 (1), Cambridge University Press, pp 5--23 (2006)
19. Ciekanski, M., Chanier, T.: Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. In: Recall, vol. 20 (2), Cambridge University Press, pp. 162--182 (2008)
20. Kay, J., Reimann, P., Yacef, K.: Mirroring of group activity to support learning as participation. In: R. Luckin, K. R. Koedinger, and J. Greer. (eds), Proceedings of AIED 2007, 13th Int. Conf. on Artificial Intelligence in Education, vol. 158, IOS Press, pp. 584--586 (2007)
21. IMS-CP: Content Package Specification (IMS consortium). (2004) http://www.imsglobal.org/content/packaging/
22. EML: Educational Modelling Language, Open University of the Netherlands (OUNL). (2000) http://www.learningnetworks.org/?q=EML
23. Koper, R.: Modelling Units of Study from a pedagogical perspective: The pedagogical metamodel behind EML. Technical Report OUNL June (2001)
24. Mce_sid: Full schema for the structured information data (instantiation component) of a Mulce corpus. (2008) http://mulce.univ-fcomte.fr/metadata/mce-schemas/mce_sid.xsd
25. Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting. In: Lagoze, C., Van de Sompel, H. (eds). Version 2.0 (2002) http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
26. Simons, G., Bird, S.: OLAC: Open Language Archives Community (2007) http://www.language-archives.org/ http://www.language-archives.org/OLAC/metadata.html
27. IMS-LD: Learning Design Specification of the IMS consortium, version 1.0, Jan (2003) http://www.imsglobal.org/learningdesign/ldv1p0/imsld_infov1p0.html
28. Betbeder, M.-L., Tissot, R., Reffay, C.: Recherche de patterns dans un corpus d'actions multimodales. In: Nodenot, T., Wallet, J., Fernandes E. (eds.) EIAH'2007 Conference: Environnements Informatiques pour l'Apprentissage Humain, Switzerland, june, pp. 533--544 (2007)
29. Calico: French national research project coordinated by E. Bruillard (ERTÉ: Technical Research Team in Education) (2008) http://calico.inrp.fr/, (French homepage).
30. Bruillard, E.: Teacher development, discussion lists and forums: issues and results. In: K. McFerrin, R. Weber, R. Carlsen, D.A. Willis (eds.). Proceedings of Society for Information Technology and Teacher Education International Conference, SITE 2008. Chesapeake, USA: AACE, pp. 2950--2955 (2008)
31. Giguet, E., Lucas, N.: Creating discussion threads graphs with Anagora. In proceedings of the 9th Computer Supported Collaborative Learning conference (CSCL'2009), Rhodes, Greece, pp. 616--620 (2009)
32. Corbel, A., Girardot, J.-J., Lund, K.: A method for capitalizing upon and synthesizing analyses of human interactions. In: W. van Diggelen & V. Scarano (eds), Workshop proceedings Exploring the potentials of networked-computing support for face-to-face collaborative learning. EC-TEL 2006, October, Crete, pp. 38--47 (2006)
33. Dyke, G., Lund, K., Girardot, J.-J.: Tatiana: an environment to support the CSCL analysis process. In proceedings of the 9th Computer Supported Collaborative Learning conference (CSCL'2009), Rhodes, Greece, pp. 58--67 (2009)
34. Tatiana: Trace Analysis tool for interaction ANAlysts, European LEAD project outcome, G. Dyke (2008). http://www.lead2learning.org/projectsite/pagina.asp?pagkey=76663
35. Apache Tomcat: implementation of Java servlet. http://tomcat.apache.org/
36. Apache Struts: a free open source framework for web applications. http://struts.apache.org/
37. Hibernate: object/relational persistence and query service. https://www.hibernate.org/
38. SiteMesh: a web-page layout system. https://sitemesh.dev.java.net/