The Proxy Model: A New Approach to Sharing and Analyzing Learning Traces Corpora

Hajer Chebil, Christophe Courtin, and Jean-Jacques Girardot

***Abstract*—Collecting and analyzing interaction traces in Technology Enhanced Learning (TEL) environments is a common practice of researchers wishing to optimize the efficiency of these environments. This paper proposes a new approach, introduced by the proxy model, to the challenge of sharing and analyzing contextualized learning traces corpora.**

***Index Terms*—Interaction trace, corpus of traces, corpus sharing, corpus analysis, learning traces, education, technology enhanced learning, proxy model.**

I. Introduction

Technology enhanced learning (TEL) environments are increasingly used in both distance and face-to-face learning situations. To optimize the efficiency of these environments, a common practice is to collect interaction traces to record learner's interactions with the TEL system. These traces can be analyzed with different targets, for example to help the tutor monitor the learner's activity, the pedagogical designer adjust the pedagogical scenario, and the researcher confirm or refute a hypothesis about the mechanisms of knowledge acquisition. The research work described here is a part of the project 1 entitled "TEL environments customization". Research groups of this project collect and analyze interaction traces to study different research questions such as the adequacy of a pedagogical scenario to a real learning situation, or the role of the awareness tools in the use of communication tools. Interaction traces are considered of a great help for a researcher who wishes to study learning acquisition processes, tools usefulness, tools efficiency, etc. These traces correspond to the recording of the use of different categories (e.g. communication) and types (e.g. forum) of computer-supported learning assistance tools. Traces are then heterogeneous because of their different models. They can also be of different natures: numeric log traces, video traces and human observations. Our work consists in proposing a platform intended for TEL researchers for sharing contextualized interaction traces corpora, and analysis tools. This platform can be seen as a benchmarking platform giving the possibility to researchers to perform comparative and cumulative analyses and to integrate their produced resources into the platform. This platform, called BEATCORP (BEnchmarking platform for Analysis of Trace CORPora), is based on the "proxy model" we propose. The research question we are going to deal with in this article is: "How to take into account the different interaction trace representations without losing semantics in a context of sharing traces corpora and analysis tools?". The remainder of this paper is organized in two main sections, the first briefly presents some approaches to the challenge of corpora and analysis tools sharing, and the second presents our proposal. Last, we conclude and describe future work.

Manuscript received May 9, 2012; revised June 14, 2012.

H. Chebil and J. J. Girardot are with the Laboratory for Information Science and Technology, Henri Fayol Institute, Ecole des Mines de Saint-Étienne, 158, cours Fauriel, 42023 Saint-Étienne Cedex 2, France (e-mail: chebil@emse.fr).

C. Courtin is with the Université de Savoie, 73376 Le Bourget du Lac Cedex.

II. Some Approaches to the Challenge

This section presents the existent approaches to sharing traces corpora and post-hoc analysis tools by researchers in the computer-supported learning field. We notice two main interrelated sharing issues: sharing corpora and sharing analysis tools of these corpora. Existent works handle this problem from the corpus sharing point of view, the tools sharing point of view, or both points of view. The simplest and easiest way to share data, and in order for this data to be understandable in a relatively reasonable time by a researcher that hasn't contributed in collecting that data, is to structure them in a particular format common to all shared data. The absence of a normalised format to represent interaction traces of computer-supported learning situations motivated researchers to propose formats that can be used to encourage sharing. The shared data are often likely to be analysed with interaction traces analysis tools. Analysis tools can be very specific and tightly related to the application domain of the learning tool or generic to assist analysis of traces collected in different learning tools. The analysis tool thus needs formatted data as input. This implies that collected traces format is either directly accepted by the analysis tool, or that these traces need to be converted into the tool input format. This need of analysing shared data consolidates the need of a standard representation format. Assuming that a consensual representation format for learning interaction traces can be proposed, analysis tools will be implemented in a way that they accept this standard format. This will increase the usability of analysis tools by facilitating the interoperability between learning environments and analysis tools. The MULCE (Multimodal Learning and teaching Corpora Exchange) project [2] focused on the importance of sharing learning and teaching corpora between researchers. This work puts emphasis on the necessity of contextualising the shared interaction data collected during collective experimental learning situations. The PSLC Datashop project [1] offers a web-based platform providing a repository of interaction traces datasets (coming from intelligent tutoring

1 http://liris.cnrs.fr/~clu-eiah/?page_id=151

systems and conforming to a particular logging model) and a suite of tools to perform exploratory analyses and visualizations on that data. The Interaction Analysis (IA) JEIRP Kaleidoscope project [4] aimed at offering a shared library of interaction analysis tools for the Technology Enhanced Learning (TEL) community. An interaction description format called "the common format" has then been proposed. The project emphasizes the complexity of proposing a common format. In fact, a trade-off has to be reached between: (1) a very generic format which enables representing a multitude of data but which may cause losses in certain data semantics, and (2) a more specific format which allows the implementation of automatic features but restrains the multitude of data to be represented. Although the initiative of this project was interesting and promising, it has not actually led to an available library of shared interaction analysis tools. The CALICO project [6] deals with sharing and analyzing discussion forums traces collected in professionalizing training sessions. A generic model of forum traces had been proposed. Once the data are expressed in the proposed format, it becomes possible to share and analyze them using the tools shared by the platform developed within the project. In this project, proposing a shared format for representing discussion forums interaction data is realistic because of the specificity of the considered interactions.

Despite the importance of this issue of data and analysis tools sharing between researchers in TEL, we notice the small number of works that have taken an interest in it. We present in the next two sections our new approach to deal with this issue of sharing. We claim that our approach is realistic in the absence of a standard format which is adopted by the TEL community. We propose a generic structure of contextualized interaction traces corpus and an approach called the "proxy approach" for sharing, querying and analyzing corpora.

III. The Proxy Approach for Sharing and Analyzing Traces Corpora

A. Trace Definition

In this section, we present what we call corpus of contextualized interaction traces. Let us first present what we mean by trace. In [8] a digital interaction trace is defined as: "a sequence of temporally situated observed elements which stems either from an interaction between humans mediated in various ways by computer, or a suite of actions and reactions between a human and a computer. Traces are possibly replayable, in which case, they become dynamic. The trace is digital because it records actions performed on computer or a digitized version of a video (showing humans in interaction or a screenshot during the interaction)" (translated from the original French definition version). As presented in [7], the concept of observed elements represents a datum relative to an observation activity. For the interpretation of the trace in a given context, a semantics is associated by means of a use model which corresponds to the notion of MTrace (trace + model).

We adopt the previous definitions while differentiating, because of their different natures, between (1) a log trace (raw or enriched) which records actions performed on a computer, and (2) a trace which is not directly interpretable by a machine and needs a human intervention to understand its content (this trace is typically an audio/video recording or a human manually collected observation). The first type of trace has a property that the second does not, which is the possibility of implementing automatic processing of the trace, to perform calculations and automatic transformations without needing human intervention. Such automatic processing needs a previous work of transcription when working on a video trace. So we will refer to these two types of traces as: trace of type I and trace of type II.

B. The BEATCORP Corpus structure

As in the MULCE project [2], we stress the need to contextualize the interaction traces using data describing the context of the observed learning situation. In fact, our aim being among others to share corpora between the researchers of the TEL community, the availability of interaction traces is not sufficient for understanding the shared data by a researcher who did not participate in the experiment. To meet this need, we suggest the integration to the corpus of any resource which permits a better understanding of the traced interactions. Generally speaking, we consider that a contextualized interaction traces corpus is composed of two main components. The first component is the corpus description which is, in turn, composed of three sub-components: (1) the general description of the corpus which gives summary information about the corpus content; (2) the description of the shared resources made available within the corpus; and (3) the description of the analytical work performed on the corpus data. This description component makes it possible to browse and query the shared corpora database by a researcher wishing to analyze one or more shared corpora. The second component is the set of the physical resources shared within the corpus. We distinguished five types of potential shared resources within a corpus: (1) pedagogical resources, which can be either (1.1) teaching-oriented, which means offered by the learning environment to the learner during his learning activity (e.g. a problem statement, a course material), or (1.2) learning-oriented, which means produced by the learner (e.g. a dissertation); (2) traces resources, which can be of type I or II (types presented above); (3) analysis resources, we distinguished three possible sub-types of an analysis resource: (3.1) imported resource, a complementary resource which is needed by the researcher to analyze the interaction traces (e.g. an interpretation model to annotate interaction events), (3.2) produced resource, a resource produced by an analysis tool used by the researcher in his analysis work, such resource makes it possible for another researcher to consult the work and eventually to enrich it, an analysis tool does not necessarily save the results of analytical work, this is due to the fact that sharing is not necessarily an objective of the researcher, and (3.3) interpretation resource, a resource produced by the researcher during his analysis work to interpret the results; (4) publication resources, any publication presenting results of a research work has to be integrated into the corpus; and (5) documentation resources

which document the corpus description (e.g. experimentation description, analysis work description).

We propose to differentiate two types of corpora: original corpora and analysis corpora. An original corpus corresponds to the observation of an experiment carried outusing a TEL environment. It is constructed by gathering resources used and collected outside the platform during the TEL experiment. It is possible that analytical work had been achieved on the corpus before its (re)construction in the BEATCORP platform, resources used and produced during the analyses must be integrated into the corpus, allowing other researchers to access, verify and enrich them [2]. In order to distinguish between an original corpus, as it was collected and described by a researcher, including analyses achieved on it outside the platform, and the analytical work performed on the corpus within the framework by means of one or several shared analysis tools, we introduce the concept of analysis corpus. A new analysis corpus is created in the BEATCORP platform in order to answer a particular research problem of a researcher or a group of researchers. Analysis can be performed on more than one shared corpus. The researcher extracts the needed data from the corpora resources that are interesting for his analysis work. We consider the possibility that a researcher reuses a resource coming from an analysis corpus previously constructed in the platform. Thus, an analysis corpus can refer either to original or analysis corpora. As far as the physical resources types to be shared within an analysis corpus are concerned, these can be of the three last resource types presented above.

C. The Proxy Model

The proxy model is seen as an intermediary layer between shared corpora and analysis tools. As already explained, our approach avoids proposing a unique model to represent interaction traces which we said will not necessarily cover all trace modelling needs. Alternatively, we propose a trace corpus' ontology. This ontology defines a set of concepts (possibly linked) that enables to describe a corpus (corpus description component) and interaction trace resources within it. Concepts composing the ontology are generic and represent data usually collected in interaction traces. The advantage of the ontology is that it can be enriched and completed according to the changing needs. Ontology concepts will be mapped with concepts retrieved in different collected interaction traces coming from different learning environments. Fig. 1 below illustrates the various layers of proxy model as well as their roles.

A corpus shared within the platform and likely to be analyzed by shared analysis tools is linked to a "corpus proxy" layer which gives access to the corpus description. This layer also defines a subset of the ontology concepts (OCS) that are identified as relevant to describe the corpus trace resource concepts. Finally, it defines instances relative to ontology concepts and the way to extract the corresponding data.

The "generic queries proxy" allows capitalizing upon the generic modules allowing the querying of the shared corpora database. The first component concerns general queries on the corpora database, it uses the corpus description defined in the "corpus proxy" layer. This description allows the definition of queries based on elements that constitute it. The second component is an engine allowing uniform querying of interaction traces shared within corpora. The inputs of this querying engine are: (1) the OCS relative to a particular corpus, (2) the concepts' instances relative to the used OCS, (3) chosen projection parameters, i.e. OCS concepts that a researcher wants to retrieve in his query results, and (4) selection conditions corresponding to filters specifying conditions on the values of some concepts of the OCS (e.g. interactions having begin date ulterior to 01/01/2012). Inputs (2) and (4) are defined in the "corpus/tool proxy" layer. This querying engine returns the query result as a list of couples (concept, value associated to it) for each interaction. The third component of this layer is a set of generic scripts for converting values returned by the querying engine. A script allows converting a value from a data type t1 to a data type t2. For example, converting a string into a date.

The "corpus/tool proxy" layer defines components specific to a particular couple (analysis tool, corpus). Indeed, it is at the level of this layer, depending on the input data format defined by the analysis tool, that are specified the OCS concepts expected in the extracted data. These extracted concepts are called projection parameters. Furthermore, it is also possible to define selection conditions (filtering) on the values of the concepts within the data to be extracted. Finally, the third component is a conversion and formatting module allowing to format data extracted from the corpus to be converted to the format expected by the analysis tool.

The "analysis tool proxy" layer allows describing the analysis tool and its expected input data format. It also includes a module for defining integration scripts of input data expected by the analysis tool. This can be useful if an analysis tool is used to analyze data coming from more than one corpus and expects a single input data stream corresponding to queried corpora. In that case, an integration script can be developed and used to integrate all extracted data in a single stream.

The "analysis corpus proxy" is related to the analysis corpus that will contain the performed analysis work. It allows keeping track of analytical work by describing it, which will allow its reproduction. This description: (1) shares metadata and general descriptions related to the analysis, (2) links a used analysis tool to one or more corpora that it analyses while specifying the used queries to extract corpus data, and (3) references used and produced resources.

The physical resources used as complementary in the analysis as well as those produced by the analysis have to be stored in the analysis corpus. The whole approach relies on standards (XML, RDF, OWL, XSchema, XQuery) and open-source tools (Protégé, eXist).

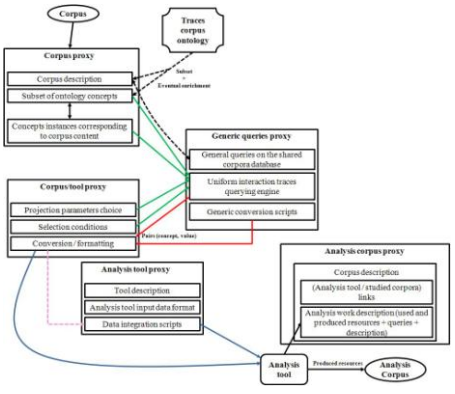

Fig. 1. Proxy model layers and their roles.

IV. Conclusion and Future Work

This paper tries to answer the research question "How to take into account the different interaction trace representations without losing semantics in a context of sharing traces corpora and analysis tools?" We proposed the proxy approach which (1) permits to share data without imposing any constraint on the data formats, (2) adds a general corpus description which is the same for all shared corpora, (3) capitalizes on generic queries which can be performed on different corpora and (4) adds specific queries and conversion methods to prepare input data needed to perform an analysis with a particular analysis tool on a particular shared corpus. Future work concerns implementing a prototype of the BEATCORP platform to validate our proxy model, and to test it by sharing an original corpus and performing analyses on it.

Acknowledgements

This work is part of the Research Cluster ISLE project (now ARC6), supported by the "Région Rhône-Alpes" in France.

References

[1] K.-R. Koedinger, K. Cunningham, A. Skogsholm, and B. Leber, "An open repository and analysis tools for fine-grained, longitudinal learner data," *Proc. of the 1st International Conference on Educational Data Mining*. Montréal, Québec, Canada. 2008, pp. 157-166.

[2] C. Reffay and M.-L. Betbeder, "Sharing corpora and tools to improve interaction analysis," *Proc. of the 4th European Conference on Technology Enhanced Learning*. Nice, France. 2009, pp. 196-210.

[3] J.-S. Sottet, G. Calvary, and J.-M. Favre, "Ingénierie de l'Interaction Homme-Machine Dirigée par les Modèles," *Premières Journées sur l'Ingénierie Dirigée par les Modèles*. Grenoble, France. 2005, pp. 67-82.

[4] A. Martínez, A. Harrer, and B. Barros, "Library of interaction analysis methods," *Deliverable of the ICALTS JEIRP project*. 2005.

[5] C. Choquet. Ingénierie et rétro-ingénierie des EIAH – l'approche REFiM. *Accreditation to supervise research*. 2007.

[6] Calico. http://woops.crashdump.net/calico/. 2012.

[7] C. Courtin and S. Talbot, "Automatic Analysis Assistant for Studies of Computer-Supported Human Interactions," *Book Chapter, Lecture Notes in Computer Science*, 5794, Learning in the Synergy of Multiple Disciplines. 2009, pp. 572-583.

[8] J.-C. Marty and A. Mille, "Analyse de traces et personnalisation des environnements informatiques pour l'apprentissage humain," *Hermès Sciences Publishing, collection IC2 informatique et systèmes d'information.* 2009.