Building and Analyzing a Corpus of Contextualized Traces Collected during a Technology Enhanced Teaching Module

Hajer Chebil, Christophe Courtin, Jean-Jacques Girardot

Building and analyzing a corpus of contextualized traces collected during a Technology Enhanced teaching module

Hajer Chebil, Christophe Courtin
Syscom Research Team
Université de Savoie
Chambéry, France
hajer.chebil@gmail.com,
Christophe.Courtin@univ-savoie.fr

Jean-Jacques Girardot
Henri Fayol Institute
Ecole des Mines de Saint-Etienne
Saint-Etienne, France
girardot@emse.fr

Abstract—**Sharing and analyzing data collected within Technology Enhanced Learning environments is an interesting issue for researchers to validate their models and systems. In this paper we present a corpus we built and analyzed in order to validate our proposed "Proxy approach" as an approach for sharing and analyzing learning data corpora.**

TEL; traces; context; corpus; sharing; analysis

I. INTRODUCTION

Digital traces collected within Technology Enhanced Learning (TEL) environments are very often used to personalize such environments in order to enhance the learning practices. Such traces are analyzed on-line to help tutors better supervise the learning sessions [1] and to adapt the activities to the learners' activity [2]. Traces can also be subject to offline analyses with different goals like reengineering of TEL environments [3], adapting a learning scenario [4], or analyzing interactions [5]. Building an authentic learning experiment is a complex and time-consuming task. As a consequence, having access to research data offers more recognition and visibility for owners [6] while it allows analyses' replication and comparison by other researchers [7]. We proposed a new approach called the "Proxy approach" [8] [9] as a base for a platform architecture that allows researchers to share learning traces corpora, without imposing a unique representation of data, and to flexibly use different analysis tools on different corpora depending on analysis needs. In this paper, we describe an experiment we performed within the Technology Institute of the University of "Savoie" during the "Object-Oriented Design / Object-Oriented Programming" (OOD-OOP) unit. Based on this experiment, we built and analyzed a corpus in order to validate our approach. The rest of this paper is organized as follows. Section II presents the pedagogical and scientific objectives that motivated us to perform this experiment. Section III gives a short overview of the "Proxy approach". Section IV describes the experiment and the learning environment. Section V presents the structure of the resulting corpus and the analyses performed on it. Section VI exposes the results followed by conclusions and a set of perspectives.

II. THE EXPERIMENT'S OBJECTIVES

The experiment we describe in this paper and that we call "OOD-OOP" (in reference to the unit) has been conducted according to two types of objectives. The first is obviously pedagogical since the experiment corresponds to an academic unit. The second is scientific and is related to the validation of our approach.

The pedagogical objective is to help students achieve a practical work of "OOD-OOP" while communicating and collaborating via a discussion forum. Using the forum aims to capitalize important information and to avoid wasting time asking a previously answered question.

The scientific objectives are (1) to build a corpus of traces collected during the experiment enriched by contextual information, and (2) to apply and validate the "Proxy approach" in order to perform analyses on the corpus using two different analysis environments.

III. OVERVIEW OF THE PROXY APPROACH

In this section, we concisely present the "Proxy approach" (see [8] and [9] for more details). This approach addresses the problems of (1) heterogeneity (semantics, formats) of collected data and the corresponding learning contexts, (2) strong coupling between data formats and analysis tools, and (3) non-reproducibility of previous analyses. Existing approaches ([5], [7], [10], [11], [12]) propose a unique representation of the data to be shared and possibly analyzed using a set of available analysis tools. In order to use the services provided by existing sharing platforms, users have to do a preliminary work of conversion to format data according to the chosen representation. However, the heterogeneity related to learning environments, pedagogical scenarios, disciplines, and analysis needs makes the definition of a standard representation very difficult. Based on this observation, we worked on a different approach which avoids imposing a yet another representation which cannot be suitable for all data representation needs. The general idea is to allow researchers to share any trace or contextual resource without having to convert the format of its content into another one. The user has only to provide a set of metadata allowing the description of the resources and their future querying. Moreover, when a user wants to query some corpus content in order to perform an analysis, s/he uses a component we call "proxy" that plays the role of an intermediary between the researcher, the corpus and the analysis tool. The “Proxy approach” relies on three models: (1) the corpus model which defines two types of corpora: (i) “initial corpus” which corresponds to all resources collected by a researcher in relation to the experiment, and optionally the resources and the descriptions relative to analyses performed outside the sharing platform which is based on the “Proxy approach”, (ii) “analysis corpus” which corresponds to analyses performed on the sharing platform using the services of the “Proxy approach” to answer a particular research question. This first model also defines the structure of a corpus which can contain different types of resources (traces, pedagogical resources, productions, etc.) and a corpus description according to the corpus description model that defines a set of describing metadata (subset of the DublinCore [13] metadata set enriched by learning/teaching-oriented metadata); (2) the semantic model, which corresponds to a taxonomy that defines a set of queryable concepts, and which is used as a semantic communication tool between researchers who share corpora and analysis tools; and (3) the operational model, which implements the operational aspect of the approach by defining several types of operations (query, convert, filter, format, fusion) that will align the semantics of the concepts defined in the semantic model with the corpora contents and the input data format of different analysis tools. The different operations are implemented using scripts. Once the different operation scripts are defined, corpora querying becomes uniform and semantic issues transparent. The three models are conceptualized as a descriptive ontology.

IV. The Experiment’s Description

The context of the experiment was a real practical class of the “OOD-OOP” module at the university with students of the multimedia department. The group was composed of nine students who have almost the same level. This work was intended to increase the students' skills in the Java object-oriented programming language. Each student had to implement a software phone (softphone) in Java with the Eclipse [14] integrated development environment. Each student could communicate with other students and with the teacher to ask for help. The teacher's role was clearly defined as catalyst and facilitator in both direct and mediated discussions. The work was planned on four sessions, but the students had to deliver their work in progress at the end of each session.

The observation we have conducted was based primarily on the use of several tools of the Moodle [15] e-learning platform. Furthermore, the two last sessions were filmed in order to keep track of synchronous conversations and what happened in class during these sessions.

During the scheduled time, the communication took place *in situ* using a software communication tool (a forum) in addition to the direct human communication (dialogue and illustrations on a classical whiteboard). The students were encouraged to participate in discussions and to post questions in the forum. The remaining time, i.e. between the sessions, work was done remotely and the students communicated with the teacher only by means of the forum. At the end of each session, they had to upload their intermediate files by means of the Moodle deposit service.

The teacher wanted to measure the degree of student/student collaboration and student/teacher collaboration. His purpose was to facilitate inquiry collaborative learning. In others words, the students would seek information from other students or the teacher knowing that the result of their investigations had to be capitalized in the forum. For this purpose, the forum was structured into several topics of discussion. These different topics dealt with different aspects of the work to be achieved by the students.

V. Corpus Structure and Analyses

After the end of the experiment, we proceeded to the construction of the “OOD-OOP” corpus. In this section, we present the different steps of the corpus construction phase. The corpus construction phase is composed of three principal steps: the first step was to collect or create the different resources that would compose the corpus; the second step was to provide a set of descriptive metadata to document the corpus and the resources which compose it; finally, the last step was to create the different scripts corresponding to the operations of the operational model and which will make the querying, extraction, and conversion of the corpus’ data possible in order to be further analyzed by two different analysis tools. In the second part of this section, we describe some analyses performed on the corpus based on the “Proxy approach”.

A. *Corpus construction*

1) *Resource collection / creation*

We distinguished different types of resources that can be shared within a corpus. Such resources can contain traces or different contextual data useful for the sharing and the understanding of the corpus. We distinguish two complementary types of traces: (1) those that are collected automatically by means of collecting modules embedded in the virtual learning environment, and (2) those that correspond to audio/video recordings of a learning session. In the case of the “OOD-OOP” corpus, we have resources corresponding to the two types of traces: (1) traces that were automatically collected within the Moodle platform and which correspond to the interactions between the students and the teacher that took place in the discussion forum, to extract these traces we executed an SQL query on the relational database hosting the Moodle data, and (2) we also have seven audio/video recordings that correspond to the two last sessions of the practical work. The other types of resources we share in the “OOD-OOP” corpus are: (1) seven pedagogical resources which correspond to resources provided by the teacher during the learning period (three course resources, a resource explaining the task to achieve, a resource containing some programming advice, a tutorial, a resource giving a solution of a part of the work done by the students); (2) forty seven production resources which correspond to resources produced by the participating learners (corresponding to intermediary and final productions of the students); (3) one documentation resource which describes the experiment which led to the corpus and (4) publication resources that correspond to the communication papers, reports or manuscripts about analyses and results that are related to the experiment and the corpus (Till now, we have one publication related to the corpus and which corresponds to the manuscript relative to the PhD work during which the described experiment was driven).

2) Corpus and resources description

The corpus description model is a part of the corpus model and contains three parts: general description of the corpus which defines general metadata about the experiment underlying the construction of the corpus (e.g. title, creator, contributor, keywords, etc.), resource description (e.g. title, subject, creator, format, etc.), and previous analysis work description (information about the date and objectives of the analysis, the used analysis tool(s), the used services that allowed data extraction, conversion and formatting). This step corresponds to the documentation of the corpus by providing a set of general metadata about the whole corpus, and the different resources that compose the corpus. It is worth noticing that descriptive metadata are not mandatory but they are very useful to contextualize the corpus content and to help researchers that don't have an idea about the experiment to understand the data without needing the help of the researchers that built the corpus. In this way, the corpus becomes self-explanatory. Furthermore, the description model allows researchers to describe in detail the previous analyses performed on the corpus which makes them reproducible. In the case of the "OOD-OOP" corpus, we do not describe analyses in the initial corpus since we used the services of the "Proxy approach" in achieving the analyses that we will describe in the next section. However, we created an analysis corpus named "ForumCollab" to study the role of the forum interactions in enhancing the student/student and student/teacher collaboration and in contributing to the capitalization of useful information.

3) Corpus querying scripts

This third step is necessary for performing analyses on the corpus resources. In fact, corpora can contain resources having different formats, and more importantly different semantics; on the other hand, analysis tools expect input data with semantics and formats that can be different from those of the corpus' original data. Defining scripts as part of the operational model allows a researcher planning to analyze some corpus data to semantically align the corpus data concepts that interest him with those of the semantic model.

In order to analyze the trace resource corresponding to forum interactions, we defined different scripts. Assuming that any digital traces resource can be converted to an XML representation, we made the choice to represent digital traces using an XML tree, and to write the scripts in XQuery. We defined an XQuery script that extracts data corresponding to forum interactions from the digital traces we exported from Moodle's database. This script allows aligning the concept "forum interaction", defined within the semantic model, with the corpus resource data. When needed, data type converting and filtering scripts can be defined. In our example, we did not need to define such scripts because original data types of the queried resources' data were compatible with our analysis needs and because we needed to extract all interactions without any filtering.

We also defined scripts to format extracted data to be analyzed using two different analysis environments. Such scripts perform the alignment between the semantic model's "forum interaction" concept and the data structure expected in the input of the analysis tools. These scripts will be described in the next section as part of the analysis phase.

B. Analyses description

To validate our approach, we used two analysis environments freely provided by researchers in TEL. The first one called Tatiana [16] is intended for researchers to achieve analyses in an iterative manner while offering different visualization tools. The second analysis environment is the CALICO platform [5] that offers a set of analysis tools allowing a better exploration of forum interactions. To analyze the extracted data using the two analysis environments we used in our approach validation, we defined one formatting script specific to each environment. The role of these scripts is to format extracted data to be ready for analysis by a specific analysis tool. So analyzing the forum interactions of the "OOD-OOP" corpus becomes simple. Once we defined scripts that extract data relative to the "forum interaction" concept defined in the semantic model, the only next remaining step consists in defining the formatting scripts that only format data, assuming that semantics are compatible between the analysis tool and the semantic model.

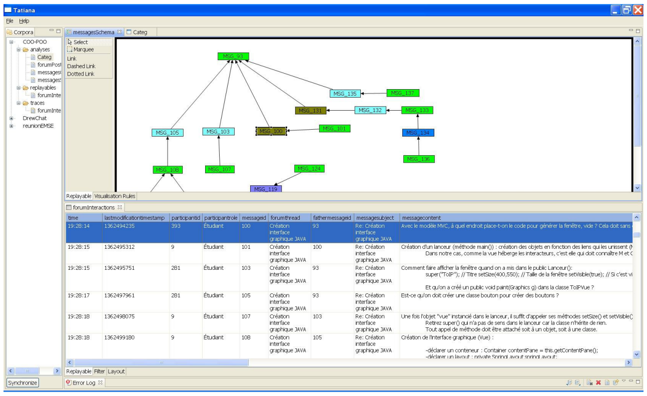
To analyze the "OOD-OOP" initial corpus data, we built the "ForumCollab" analysis corpus to study the research question "to what extent did the use of the discussion forum improve the student/student and student/teacher collaboration, and contribute to the capitalization of information?". "ForumCollab" is composed of a set of physical resources and a description. The resources are as follows: one documentation resource which describes the analyses performed within the corpus and the studied research question, and eleven analysis resources among which eight are produced by the Tatiana analysis tools, two resources correspond to snapshots of the CALICO tools interfaces, and an interpretation resource which document our understanding of the results. The description of the corpus is composed of three parts: (1) a general description of the corpus using a set of metadata (e.g. creator, contributor, creation date, objectives, keywords, etc.); (2) a physical resources description using a set of metadata (e.g. title, creation date, format, type, producing tool, etc.); and (3) descriptions of the analyses performed within the corpus (e. g. begin and end dates, creators, objectives, reference to the used analysis tools, reference to the produced analysis resources, etc.). Sharing the produced resources and the descriptions of the performed analyses allows understanding and reproducing those analyses, which can be very useful for researchers.

We will now describe the analyses we performed using Tatiana and CALICO to try to answer the studied research question by analyzing the collected forum interaction traces. We performed two analyses using Tatiana and a third using CALICO.

After importing forum interactions into Tatiana, we used the graphical representation (cf. Figure 1.) to visualize interactions as a graph in which colored vertices correspond to messages posted in the forum. Each color corresponds to a participant, and an arrow between two vertices expresses the "reply to" relation. This graphical representation of the forum interactions showed us that almost all the interactions took place between the teacher and the students. The students usually reply to the first message, posted by the teacher to start the discussion thread, to ask their questions. And the teacher was the only participant who answered the students' questions. Figure 1. below illustrates the Tatiana interface with a tabular representation of the interactions synchronized with the graphical one. Synchronizing the two representations provides access to the contents of posted messages by simply selecting a box in the graphical representation. In the second analysis, we focused on the semantics of the posted messages and we worked on a categorization of the forum posts. Each category is represented by a different color. The defined categories represent what exchanged messages represent with respect to the learning activity. We distinguished thirteen categories: (1) ask a question; (2) re-ask a previously asked question; (3) answer a question by giving an explanation; (4) answer a question by giving an example; (5) answer a question by giving an explanation and an example; (6) contest an answer; (7) comment an answer; (8) answer one's own question; (9) ask and answer a question; (10) provide some information; (11) start a discussion thread to treat some subject; (12) start a discussion thread to give some information; and (13) start a discussion thread to ask a question. In defining these categories, we tried to think about different post categories we can find in such a discussion forum. All these categories weren't found in the collected forum interactions. We don't claim that these categories are exhaustive and new categories can be added when needed. Figure 2. illustrates a graphic representation of all the messages of the forum colored according to their type categories. This allows having a visual representation of the different categories of the exchanged messages and their proportions. In the graph of Figure 2. , we can notice that most of the exchanged messages corresponded to asking questions (shades of blue) and answering questions (shades of green).

The CALICO platform offers a set of tools that enhance reading forum interactions. Some tools offer a quantitative view of data, when others are intended to perform in depth analyses of the message contents. Figure 3. illustrates the use of the concordancer named Concordagora to search for two terms specific to questioning ("*comment*" (how …?) and "*Est-ce que*" (is ...?)) and that are often used to ask questions. The use of this tool has helped us to have a better and faster reading of the interactions.

The analyses presented in this section enabled us to study the research question we are interested in, in relation with the role of technology in enhancing collaboration and information capitalization. We noticed that interaction was essentially student/teacher because the students asked questions and the teacher was the only participant who answered them. The objective of enhancing collaboration in the forum was so partly reached and more work has to be done to encourage students to collaborate more- (e.g. giving extra marks to students that are more active on the forum or students that give a correct answer within a given period of time). The second objective which is to capitalize useful information to avoid asking the same question several times, was achieved. Students searched for existing discussions before initiating new ones.



time	lastmodificationtimestamp	participantid	participantrole	messageid	forumthread	fathermessageid	messagesubject	messagecontent
19:28:14	1362494235	393	Étudiant	100	Création interface graphique JAVA	93	Re: Création interface graphique JAVA	Avec le modèle MVC, à quel endroit place-t-on le code pour générer la fenêtre, vide ? Cela doit sans
19:28:15	1362495312	9	Étudiant	101	Création interface graphique JAVA	100	Re: Création interface graphique JAVA	Création d'un lanceur (méthode main()) : création des objets en fonction des liens qui les unissent (M Dans notre cas, comme la vue héberge les interacteurs, c'est elle qui doit connaître M et C

19:28:15	1362495751	281	Étudiant	103	Création interface graphique JAVA	93	Re: Création interface graphique JAVA	Comment faire afficher la fenêtre quand on a mis dans le public Lanceur():
								super("ToIP"); // Titre setSize(400,550); // Taille de la fenêtre setVisible(true); // Si c'est vi
								Et qu'on a créé un public void paint(Graphics g) dans la classe ToIPVue ?

19:28:17	1362497961	281	Étudiant	105	Création interface graphique JAVA	93	Re: Création interface graphique JAVA	Est-ce qu'on doit créer une classe bouton pour créer des boutons ?
19:28:18	1362498075	9	Étudiant	107	Création interface graphique JAVA	103	Re: Création interface graphique JAVA	Une fois l'objet "vue" instancié dans le lanceur, il suffit d'appeler ses méthodes setSize() et setVisible(Retirez super() qui n'a pas de sens dans le lanceur car la classe n'hérite de rien. Tout appel de méthode doit être attaché soit à un objet, soit à une classe.
19:28:18	1362499180	9	Étudiant	108	Création interface graphique JAVA	105	Re: Création interface graphique JAVA	Création de l'interface graphique (Vue) : -déclarer un conteneur : Container contentPane = this.getContentPane(); -déclarer un layout : private SpringLayout springLayout;

Figure 1. Snapshot of the Tatiana analysis tool interface allowing the synchronization of graphical and tabular representations of the forum interactions

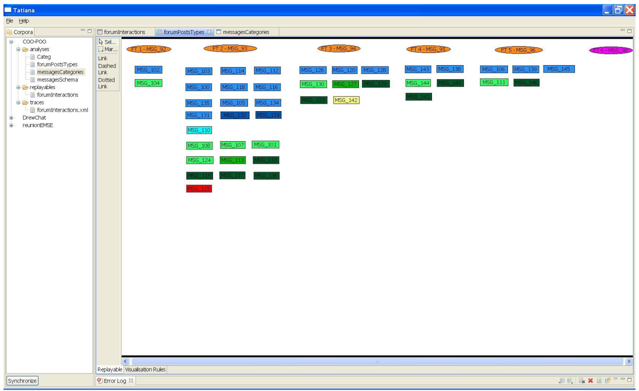
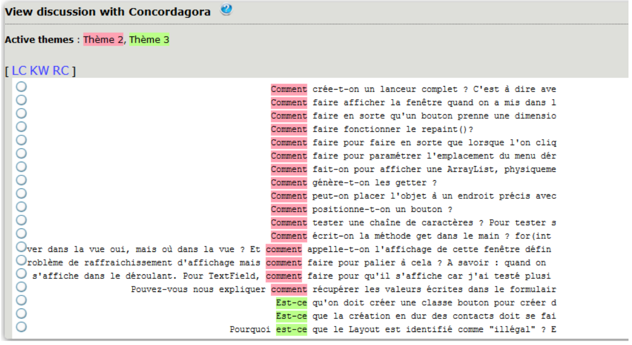


Figure 2. Snapshot of the Tatiana interface of a graph allowing the visualization of the different messages exchanged in the discussion forum colored according to their categories



View discussion with Concordagora

Active themes : Thème 2, Thème 3

[LC KW RC]

```
○                                     Comment crée-t-on un lanceur complet ? C'est à dire ave
○                                     Comment faire afficher la fenêtre quand on a mis dans l
○                                     Comment faire en sorte qu'un bouton prenne une dimensio
○                                     Comment faire fonctionner le repaint()?
○                                     Comment faire pour faire en sorte que lorsque l'on cliq
○                                     Comment faire pour paramétrer l'emplacement du menu dér
○                                     Comment fait-on pour afficher une ArrayList, physiqueme
○                                     Comment génère-t-on les getter ?
○                                     Comment peut-on placer l'objet à un endroit précis avec
○                                     Comment positionne-t-on un bouton ?
○                                     Comment tester une chaîne de caractères ? Pour tester s
○                                     Comment écrit-on la méthode get dans le main ? for(int
○ver dans la vue oui, mais où dans la vue ? Et comment appelle-t-on l'affichage de cette fenêtre défin
○roblème de raffraichissement d'affichage mais comment faire pour palier à cela ? A savoir : quand on
○ s'affiche dans le déroulant. Pour TextField, comment faire pour qu'il s'affiche car j'ai testé plusi
○                   Pouvez-vous nous expliquer comment récupérer les valeurs écrites dans le formulair
○                                     Est-ce qu'on doit créer une classe bouton pour créer d
○                                     Est-ce que la création en dur des contacts doit se fai
○                           Pourquoi est-ce que le Layout est identifié comme "illégal" ? E
```

Figure 3. Snapshot of the interface of the Concordagora tool (CALICO platform) used to analyze the forum interactions

VI. Results

As we already explained, carrying out the experiment and building the corpus was essentially planned to demonstrate that the "Proxy approach" is an interesting alternative to existing approaches regarding the issues of sharing and analyzing corpora in a technology enhanced learning context. In fact, avoiding imposing a new representation that has to be adopted to be able to share and analyze data guarantees a better acceptance of the approach. Documenting the corpus with a set of metadata is the most important effort required from the researcher. This work will however be very rewarding because the corpus will be more visible and can be used by other researchers. To be able to query the corpus and to analyze it using an analysis tool, researchers have to define some scripts which will align the concept(s) defined in the semantic model and that are interesting to study, convert, filter and format data. Furthermore, this approach is incremental and participative and a researcher who shares a new corpus isn't asked to define all the scripts necessary for querying that corpus. A researcher only defines the scripts that are needed for an analysis he wishes to perform. The "Proxy approach" not only presents the advantage that researchers can share data and analysis tools without needing to invest an important effort, but also proposes a solution to the heterogeneity problems. In fact, this approach being incremental, the semantic model can be enriched to add new concepts needed for new analyses.

The analyses we performed on the "OOD-OOP" corpus allowed us to validate our approach by using two different analysis tools to perform different analyses on the same corpus. The two analysis environments are different and have been designed by different research teams and for different purposes. The analysis results are shared within the resulting analysis corpus and can be reused by other researchers.

VII. Conclusions and Perspectives

In this paper, we presented an experiment that allowed us to build a corpus and to analyze it in order to study the validity of our "Proxy approach". The "Proxy approach" is incremental and participative and presents an interesting alternative to existing approaches. This approach is based on three models: the corpus model, the semantic model and the operational model. We presented the different steps we needed to build the corpus and the analyses we performed on the corpus.

Future work will concern giving a public access to the corpora we built and to the scripts we defined to query it. Furthermore, some interesting processing can be developed in order to automatically extract documentation data and thus minimize the researcher effort. Another interesting perspective is to develop a script generator that uses the semantic model with a synonym dictionary to generate scripts that can be validated or not by researchers.

Acknowledgment

Special thanks go to the "Rhône-Alpes" French region for supporting this work as part of the Research Cluster ISLE project (now ARC6).

References

[1] V. Guéraud, J. M. Adam, J. P. Pernin, G. Calvary, and J. P., David, "L'exploitation d'Objets Pédagogiques Interactifs à distance: Le projet Formid". Sticef, vol. 11, pp. 109–164, 2004.

[2] S. Nogry, S. Jean-Daubias, and N. Guin, "How to combine objectives and methods of evaluation in iterative ILE design: lessons learned from designing Ambre-add," Interactive Learning Environments, vol. 20(2), 2012, pp. 155–175.

[3] C. Choquet and S. Iksal, "Modeling tracks for the Model-driven re-engineering of a TEL system," Journal of Interactive Learning Research, vol. 18(2), 2007, pp. 161–184.

[4] C. Ferraris, L. Vignollet, C. Martel, A. Harrer, and Y. A. Dimitriadis, "Competitive challenge on adapting activities modeled by CSCL scripts," Proc. 8th International Conference on Computer Supported Collaborative Learning, vol. 2, 2009, pp. 57-64.

[5] E. Giguet, N. Lucas, F. M. Blondel, and E. Bruillard, "Share and explore discussion forum objects on the Calico website," Proc. 8th International Conference on Computer Supported Collaborative Learning, 2009, pp. 616-620.

[6] G. King, "An introduction to the Dataverse network as an infrastructure for data sharing," Sociological Methods and Research, vol. 36(2), 2007, pp. 173-199.

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

[8] H. Chebil, C. Courtin, and J. J. Girardot, "The Proxy model: a new approach to sharing and analyzing learning traces corpora", International Journal of Information and Education Technology, vol. 2(4), 2012, pp. 208-211.

[9] H. Chebil, C. Courtin, and J. J. Girardot, "An ontology-based approach for sharing and analyzing learning trace corpora," Proc. IEEE Sixth International Conference on Semantic Computing, 2012, pp. 101-108.

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

[11] D. Bouhineau, V. Lunego, and N. Mandran, "Open platform to model and capture experimental data in technology enhanced learning systems", Workshop Data Analysis and Interpretation for Learning Environments, 2013.

[12] V. Butoianu, P. Vidal, K. Verbert, E. Duval, and J. Broisin, "User context and personalized learning: a federation of contextualized attention metadata," Journal of Universal Computer Science, John Wiley and Sons, vol. 16(16), pp. 2252-2271.

[13] DCMI, Dublin Core Metadata Initiative, 1994, http://dublincore.org/

[14] Eclipse, Eclipse Integrated Development Environment, http://www.eclipse.org/

[15] Moodle. Moodle Learning Management System: https://moodle.org/

[16] G. Dyke, K. Lund, J.J. Girardot, "Tatiana: an environment to support the CSCL analysis process," Proc. International Conference on Computer Supported Collaborative Learning, 2009, pp. 58-67